LEVERAGING TARGETED MARKETING DATA IN TRAVEL DEMAND MODELING: VALIDATION AND APPLICATIONS

A Dissertation
Presented to
The Academic Faculty

by

Josephine D. Kressner

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Civil and Environmental Engineering

Georgia Institute of Technology
May 2014

LEVERAGING TARGETED MARKETING DATA IN TRAVEL DEMAND MODELING: VALIDATION AND APPLICATIONS

Approved by:

Dr. Laurie A. Garrow, Committee Chair
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Steven P. French
Dean of the College of Architecture
Georgia Institute of Technology

Dr. Jeffrey Newman
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Catherine Ross
School of City and Regional Planning
Georgia Institute of Technology

Dr. Kari E. Watkins
School of Civil and Environmental Engineering
Georgia Institute of Technology

Date Approved: April 3, 2014

To Abbie and Bryan, for their endless support.

ACKNOWLEDGEMENTS

I wish to thank my committee members who were more than generous with their expertise and precious time: Dr. Laurie Garrow for starting me on this path, Dean Steve French for providing support at the times it was needed most, Dr. Jeff Newman for your clear-cut feedback, Dr. Catherine Ross for supporting me from the start to the end, and Dr. Kari Watkins for being a pillar of strength and a priceless mentor that I have trusted. I would also like to thank Dr. Rick Donnelly, who has been incredibly helpful, invaluable for maintaining a balanced perspective, and who has inspired the fortitude to complete this dissertation.

My other champions include Greg Macfarlane and Candace Brakewood, who are incredible individuals, both academically and in character. May we forever be bonded through the trials and tribulations of doctoral candidacy.

To both Abbie and Bryan, your daily support has been, and will always be, incredibly personally sustaining. To my parents, Tim and Denise, I am who I am because of you. And to Tamie and Lyndsay, thanks for playing school with me twenty-some odd years ago over and over again.

This work is supported in part by the National Science Foundation Graduate Research Fellowship Program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

To date, the collection of comprehensive household travel data has been a challenge for most metropolitan planning organizations (MPOs) and state departments of transportation (DOTs) due mainly to high costs. Urban population growth, the expansion of metropolitan regions, and the general unwillingness of the public to complete surveys conflict with limited public funds. The purpose of this research is to leverage targeted marketing data, sometimes referred to as consumer data or just simply marketing data, for travel demand modeling applications. This research reveals a first step in exploring the use of targeted marketing data for representing population characteristics of a region.

Four studies were completed: an aggregate validation, a household-level validation for hard-to-reach population groups, an airport passenger model, and a residential location choice model. The two validation studies of this work suggest that targeted marketing data are similar to U.S. Census data at small geographic levels for basic demographic and socioeconomic information. The studies also suggest that the existing coverage errors are at least similar, if not lower than, the levels of those in household travel surveys used today to build travel demand models. The two application studies of this work highlight the benefits of the targeted marketing data over traditional household travel surveys and U.S. Census data particularly well, including the additional behavioral information available at the household-level and the very large sample sizes.

These results suggest that the combination of targeted marketing data with other third-party and non-traditional data could be particularly powerful. It offers tremendous opportunities to enhance, or even transform, existing travel demand modeling systems and data collection practices. Inexpensive, up-to-date, and detailed data would allow researchers and decision-makers alike to better understand travel behavior and to be more equipped to make important transportation-related decisions that affect our lives each day.

CHAPTER I

INTRODUCTION

1.1 Background and Motivation

To date, the collection of comprehensive household travel data has been a challenge for most metropolitan planning organizations (MPOs) and state departments of transportation (DOTs). The primary limiting factor is the high cost of survey-based data collection compared to most other planning functions. Even the best household travel surveys conducted today, including GPS-enabled surveys, face declining sample sizes and response rates. Researchers and practitioners continually make improvements [15, 16], but ultimately budgets have limited these steps. In 2011, the MPO of the Atlanta region conducted a survey that cost $2 million and represented just 0.5% of the region's households with only a 5.9% response rate [12, 13]. Likewise, from 2010-2012 California's DOT completed a state-wide survey that cost just over $10 million, representing only 0.4% of the households with a response rate of 2.0% [1, 14].[1] Typically, these surveys will not be collected for at least another ten years. Even the minimal costs associated with surveys for the smallest MPOs will tend to exceed the annual budget of the MPO for all planning purposes, which often results in the decision to not collect data at all [5].

Urban population growth, the expansion of metropolitan areas, and the general unwillingness of the public to complete surveys conflict with limited public funds. Despite these factors, the data from episodic household travel surveys are still used to build large portions of traditional travel demand models, which are used to identify transportation infrastructure and policy projects that achieve important regional goals. Correspondingly, a majority of federal, state, regional, and local transportation funding decisions are based on inadequate, and often outdated data.

[1] Response rates lower than about 90% are likely to produce severe nonresponse biases, and results are considered to be seriously flawed [5].

Even more, transportation planners have become attuned to the sensitivity of urban models, particularly with activity-based models [4, 10, 11]. The validity and reliability of the demographic and socioeconomic inputs of these models are important [2], particularly when analyzing the model outputs at a disaggregate level of detail. Further, legislation in MAP-21 could lead to the need for more detailed travel demand modeling outputs, causing them to be scrutinized more and more closely. Many MPOs and DOTs find it difficult to pay for the up-to-date, detailed, and disaggregate data required to build the travel demand models that are desired by federal decision-makers today [3]. Accordingly, the question naturally arises as to whether more accurate and up-to-date travel demand models using non-traditional data can be developed.

Meanwhile, we are living in a computer-driven world that is inundated with data. Third-party data are inexpensive, prolific, and information-rich. They offer tremendous opportunities to discover new information. Even more, those data collected by third-parties offer huge cost savings over traditional survey-based data collection, which could potentially fulfill the wide-reaching need for affordable, representative household data. Targeted marketing data, which make up one type of third-party data available today, provide a detailed and current picture of the nation's population. These data are typically used commercially for advertising purposes to specific markets, tracking relevant information like home addresses, demographics, socioeconomics, housing type and ownership, vehicle ownership, occupation, lifestyle classification, behavioral preferences, and hobbies. There is an opportunity to leverage these sociodemographic data for travel demand modeling applications. Because the data are current, detailed, and relatively inexpensive, there is the potential to keep regional travel demand models in sync with population trends, movements, and patterns, allowing transportation planning and research questions to be explored with models that are able to capture finer levels of detail.

1.2 Research Objectives

Most transportation planners will point to two main concerns with using third-party targeted marketing data in travel demand modeling: (1) the level of accuracy and representativeness of the data are unknown, and (2) targeted marketing firms use propriety algorithms to populate many of the variables, and therefore the integrity of the data is difficult to verify. Without knowing the method by which the data are obtained, imputed, and cleaned, does the data still provide value? To address these concerns and related questions, this work is organized into two main research objectives.

The first main research objective is to validate targeted marketing data by studying its representativeness of the population. This objective is studied at two different levels: an aggregate level and a household-level. The data are compared to U.S. Census data and a recent household travel survey for several different variables often used in travel demand modeling at the Census block group or tract level, depending on the variable. This aggregate study determines how the sample sizes, coverage rates, and coverage errors compare to the other datasets. At the household-level, the accuracy of the data is examined for several selected neighborhoods dominated by populations that are historically difficult to reach or have very low survey response rates.[2] Historically hard-to-reach populations in transportation related surveying include lower income households, large families, and zero-vehicle households. By comparing self-reported socioeconomic information to that recorded in targeted marketing data, a measure of accuracy for the targeted marketing data can be estimated.

The second main research objective is to test the usability and effectiveness of targeted marketing data. This objective is explored with two simple applications related to travel demand modeling. One of the applications determines if it is feasible to use targeted marketing data to model the location of non-airport trips-ends throughout a region for home-based airport trips. It also tests if lifestyle clustering variables from targeted marketing firms are

[2]The neighborhoods were selected for a stated preference transit study. The IRB approved survey allowed us to directly link the targeted marketing data to the survey data house by house. We could not identify another recent data source in the Atlanta region that was linkable at the household-level for this comparison.

useful for predicting travel behavior. This particular application is a part of the airport passenger model of the Atlanta Regional Commission's four-step travel demand model. It is a sub-model that could clearly and simply be separated from the rest of the model for direct comparison. Notably, the Atlanta region was the first to incorporate a separate airport trip model into their four-step travel demand model due to the fact that more airport-related trips are made in this region than in most other regions. Another application further tests the prediction accuracy of lifestyle clustering variables, and it also determines if targeted marketing data in combination with other third-party data can be used to effectively model residential location choices, which significantly affect travel decisions.

1.3 Major Contributions

This work makes four major contributions. Most importantly, this work provides directional evidence that targeted marketing data improves model fit significantly, indicting that targeted marketing data is worth further consideration. The large sample size and the additional behavioral preference information available at the household-level make the data highly predictive of both short-term travel decisions, like trips to the airport, and long-term travel decisions, like residential location decisions. In the case of the airport distribution model, the adjusted R-squared value increased from 0.2773 to 0.4635 with the addition of lifestyle clustering variables to the base model that represents the model currently used by the Atlanta Regional Commission. This amounted to an increase of 67.1% of explanatory power by the model. In the case of the residential location choice model, the large sample size of the targeted marketing data allowed for a robust model to be specified, which was demonstrated through Monte Carlo experiments. The parameter estimates associated with the lifestyle clustering variables were also more significant than any of the other variables when considering their combined effect.

By improving the predictive ability of parts of a travel demand model, transportation planners such as those from MPOs and DOTs are able to more effectively identify and fund transportation infrastructure and policy investments that achieve important regional goals like economic growth and pollution reduction at the federal, state, regional, and local

levels. The impact of investment and policy decisions made with travel demand models include decreasing congestion on a specific corridor to economic growth of a whole region, improving road safety, and attainment of the Environmental Protection Agency's (EPA) air quality standards.

Furthermore, if the sociodemographic information available in targeted marketing data at the household-level is used, a travel demand model can be built with smaller, more detailed traffic analysis zones. For example, the research undertaken in this dissertation led the Atlanta Regional Commission to purchase targeted marketing income data for use in their activity based model this past year. Due to the fact that Census income data was not available at a low enough level of geographic aggregation for their new model, the more dissaggregate targeted marketing income data proved very useful. The data's low cost make it very viable as a supplemental data source.

A second major contribution is with respect to sample sizes. Monte Carlo simulations are run for the residential location choice model, varying the number of households and the number of random alternatives included in the model. Results demonstrate the importance of larger sample sizes by visualizing the variability of model estimates when using smaller samples. Household travel surveys typically represent less than 1.0% of the population with around 2,500 to 10,000 households depending on the size of the region. Because of the prohibitively high cost of household travel surveys, larger sample sizes are no longer feasible. Targeted marketing data therefore provide an affordable alternative for obtaining adequate sample sizes in cases where detailed demographic and socioeconomic information are needed in modeling.

A third major contribution relates to the representativeness of the targeted marketing data at an aggregate level. Results show that TM data are similar to U.S. Census data at the aggregate level, particularly for age, gender, household income, and the presence of children. The largest discrepancies are associated with educational attainment and ethnicity, which is likely due to the fact that these variables are imputed more than other variables. However, these discrepancies are comparable to those observed in the household travel survey. This suggests that the techniques currently used for correcting biases in survey-based data could

be applied in a similar way to the targeted marketing data to produce unbiased data at an aggregate level.

A final major contribution showed that in the worst-case scenario targeted marketing data match self-reported data for hard-to-reach populations at rates ranging from 17.4% to 94.5% depending on the variable. The self-reported data show that incorrect household-level data randomly occur across all populations in relation to age, gender, household income, number of adults in the household, and housing tenure. It does not randomly occur across ethnicity or marital status groups. This indicates that particular sampling adjustments and weighting will need to be utilized to correct the data for hard-to-reach groups particularly when using the data at the household-level.

1.4 Research Context

Targeted marketing data lack real trip information (origin, destination, and route choice of trips). This lack of trip data makes the use of targeted marketing data limited in a travel demand modeling context. The data must still be used in combination with some other source of trip information, whether that be a household travel survey, a specialized survey like the airport intercept survey used in the first application tested in this work, or other emerging GPS or cell phone data sources in the future.

However, a wide range of applications do exist for which the targeted marketing data alone are particularly suited. Most straightforwardly, the inexpensive household-level data offer modelers the ability to work at lower levels of geographic aggregation than is possible with demographic data from the U.S. Census. For example, the Atlanta Regional Commission recently purchased targeted marketing income data for use in the ongoing development of the region's activity based model, as previous stated. Secondly, the targeted marketing data provide a unique opportunity to look at research questions related to travel behavior and travel demand modeling that cannot be answered with data traditionally used in the transportation field. The behavioral and preference data available in targeted marketing data are strong predictors of how individuals make choices, both long- and short-term. Many questions that are very relevant to transportation modelers can be examined in new

ways using this data (e.g., vehicle ownership, residential location choice, long-distance travel tendencies, mobility and other ailments, etc.).

Due to the promising results of the work in this dissertation, future research will look for ways to incorporate trip-making behavior with targeted marketing data. If this is successful, a combination of targeted marketing data with other data could potentially replace traditional household travel surveys altogether. In the meantime though, targeted marketing data are viable as supplemental data due to their extremely low cost and the vast amount of information available.

1.5 Dissertation Structure

Each of the following chapters of this work is in journal format. Each chapter begins with an abstract. This is followed by background and motivation for the particular research in that chapter, a discussion of the methodologies used, and a conclusion that summarizes the main findings, including suggestions for future research. Additionally, each chapter has a separate list of referenced literature.

Chapters 2 and 3 include the work related to validation. Chapter 2 completes the aggregate comparison of targeted marketing data to U.S. Census data and a household travel survey. The comparison is conducted with frequency distributions for eleven different variables that are relevant to travel demand modeling. This chapter was accepted for poster presentation at the 2014 Annual Meeting of the Transportation Research Board. It will be published in the *Transportation Research Record* [7]. Chapter 3 reports on the pairwise household-level comparison between targeted marketing data and self-reported survey data for primarily hard-to-reach population groups. This disaggregate comparison also reports on the randomness of the incorrect data. This chapter was accepted for poster presentation at the 2014 Annual Meeting of the Transportation Research Board. It will be published in the conference proceedings [9].

Chapters 4 and 5 include the work related to applications. Chapter 4 describes the results from using targeted marketing data with the associated lifestyle clusters in an airport passenger model. This chapter was published in *Transportation Research Record* as a part of

the Airport Cooperative Research Program's Graduate Research Award Program on Public-Sector Aviation Issues [6]. Chapter 5 investigates the use of targeted marketing data in a basic residential location choice model. It also tests model variability with varying numbers of observations and alternatives included in the model using Monte Carlo simulations. This chapter will be submitted for publication consideration [8]. The final chapter summarizes the results and suggested future research from Chapters 2 through 5.

1.6 References

[1] California Department of Transportation. "2010-2012 California Household Travel Survey Final Report." `http://www.dot.ca.gov/hq/tsip/otfa/tab/documents/chts_finalreport/FinalReport.pdf`, 2013. Accessed on 14 November 2013.

[2] Cambridge Systematics, Inc. "Travel Model Validation Practices Peer Exchange White Paper." Technical report, Federal Highway Administration, 2008.

[3] Cambridge Systematics, Inc., NuStats, N. McGuckin, and E. Ruiter. "NCHRP Report 588: A guideboook for using American Community Survey data for transportation planning." Technical report, Transportation Research Board of the National Academies, Washington, D.C., 2007.

[4] J. Castiglione, J. Freedman, and M. Bradley. "Systematic Investigation of Variability due to Random Simulation Error in an Activity-Based Microsimulation Forecasting Model." *Transportation Research Record: Journal of the Transportation Research Board*, 1831:76–88, 2003.

[5] S. P. Greaves and P. R. Stopher. "Creating a synthetic household travel and activity survey: Rationale and feasibility analysis." *Transportation Research Record: Journal of the Transportation Research Board*, 1706:82–91, 2000.

[6] J. D. Kressner and L. A. Garrow. "Lifestyle Segmentation Variables as Predictors of Home-Based Trips for Atlanta, Georgia Airport." *Transportation Research Record: Journal of the Transportation Research Board*, 2266:20–30, 2012.

[7] J. D. Kressner and L. A. Garrow. "Using Third-Party Data for Travel Demand Modeling: A Comparison of Targeted Marketing, Census, and Household Travel Survey Data." *Transportation Research Record: Journal of the Transportation Research Board*, 2014. Accepted.

[8] J. D. Kressner and L. A. Garrow. "Leveraging targeted marketing data in travel demand modeling: An application in residential location choice modeling.". Georgia Institute of Technology. Working paper, 2014.

[9] J. D. Kressner, M. F. Carragher, and K. E. Watkins. "A Household-Level Pairwise Comparison of Targeted Marketing Data and Self-Reported Survey Data." In *Proceedings of the 2014 Annual Meeting of the Transportation Research Board*, 2014.

[10] J. D. Lemp, L. B. McWethy, and K. M. Kockelman. "From Aggregate Methods to Microsimulation: Assessing Benefits of Microscopic Activity-Based Models of Travel Demand." *Transportation Research Record: Journal of the Transportation Research Board*, 1994:80–88, 2007.

[11] R. M. Pendyala and C. R. Bhat. "Validation and Assessment of Activity-Based Travel Demand Modeling Systems." In *Innovations in Travel Demand Modeling Conference, Transportation Research Board*, 2006.

[12] PTV NuStats. "Regional Travel Survey: Final Report." Technical report, Atlanta Regional Commission, 2011.

[13] G. Rousseau. "Atlanta Regional Commission Transportation Coordinating Committee Meeting, 8 October 2010." 2010.

[14] State of California. "Strategic Growth Plan: Bond Accountability: California Household Travel Survey." `http://bondaccountability.resources.ca.gov/Project.aspx?ProjectPK=0540-OCA09017-4&pid=4`, 2013. Accessed on 22 July 2013.

[15] P. R. Stopher and S. P. Greaves. "Household travel surveys: Where are we going?." *Transportation Research Part A: Policy and Practice*, 41:367–381, 2007.

[16] P. R. Stopher, R. Alsnih, C. G. Wilmot, C. Stecher, J. Pratt, J. Zmud, W. Mix, M. Freedman, K. Axhausen, M. Lee-Gosselin, A. E. Pisarski, and W. Brog. "NCHRP Report 571: Standardized procedures for personal travel surveys." Technical report, Transportation Research Board of the National Academies, Washington, D.C., 2008.

CHAPTER II

AGGREGATE VALIDATION

J. D. Kressner and L. A. Garrow. "Using Third-Party Data for Travel Demand Modeling: A Comparison of Targeted Marketing, Census, and Household Travel Survey Data." *Transportation Research Record: Journal of the Transportation Research Board*, 2014. Accepted

2.1 Abstract

This research investigates how targeted marketing (TM) data can be used as a source for up-to-date demographic and socioeconomic information. TM data provide several advantages over U.S. Census data, including the ability to incorporate additional behavioral information through lifestyle variables and conduct longitudinal studies at a low cost. We describe TM data and compare an Atlanta, Georgia sample to Census data. In parallel, we also compare the most recent household travel survey conducted by the Atlanta Regional Commission to Census data using both weighted and unweighted survey data. Results show that the distributions of sociodemographic variables are similar, particularly for age, gender, household income, and the presence of children. The largest discrepancies between the TM and Census data are associated with educational attainment and ethnicity; however, these discrepancies were comparable to those observed in the household travel survey.

2.2 Introduction

In recent years, there has been increasing interest in using non-traditional data sources for travel demand modeling applications. The interest is motivated in part by the explosion of large, third-party data sources. These big datasets, which range from mobile phone signal traces and global positioning system (GPS) data to transit smart card or credit card spending patterns, collectively provide detailed spatial and temporal data about individuals' behaviors and mobility patterns, often in real-time. The data provide metropolitan

planning organizations (MPOs) the opportunity to collect detailed and up-to-date information about its residents, non-residents, and commercial users, often at a fraction of the cost of traditional household travel surveys and commercial vehicle surveys.

Today, many MPOs spend millions of dollars on episodic household travel surveys. In 2011, the Atlanta Regional Commission (ARC) conducted a travel survey of 10,278 households (representing 0.5% of the population in the 20-county Atlanta region) at a cost of $2 million, or approximately $200 per completed survey [20]. This survey had a final response rate of only 5.93% [19]. From 2010-2012, the California Department of Transportation (Caltrans) completed a travel survey of 42,431 households at a cost of just over $10 million, or approximately $235 per completed survey with a final response rate of 2.0% [4, 21]. These types of household surveys currently form the backbone of travel demand modeling systems. As stated by the ARC, the purpose of its household travel survey is to "improve the ARC travel demand forecasts, in both its aggregate four-step trip-based model and its disaggregate activity-based model" [19]. Similarly, Caltrans notes that its California Household Travel Survey (CHTS) will be "used for the statewide model and regional travel models [and that the] CHTS data will be used to develop and calibrate regional travel demand models to forecast the 2015, 2020, 2035, and 2040 greenhouse gas emissions (GHG) and enable Senate Bill 375 and Senate Bill 391 implementation" [5].

Given that travel demand models are used to evaluate a wide range of transportation policies, the question naturally arises as to whether we can develop more accurate and up-to-date travel demand forecasting models using non-traditional data, either as a supplement to or eventual replacement for household travel surveys. However, before third-party data can be integrated into transportation applications, the accuracy and representativeness of these data need to be evaluated.

In this paper, we investigate the representativeness of targeted marketing (TM) data by comparing it to U.S. Census data for several different commonly used variables in travel demand modeling. In parallel, we compare a household travel survey (HHTS) to Census data. We also provide examples of how TM data can be linked with other types of transportation data. The remainder of this paper is presented in several sections. First, the

TM data are described and the benefits and limitations of using these data for transportation applications are discussed. Next, the coverage and representativeness of the TM data are assessed through comparisons between it and the 2011 ARC weighted and unweighted HHTS to Census data. The next sections provide a review of applications that have used TM data; some of these examples provide insights into how privacy and confidentiality concerns can be addressed by researchers when combining TM and transportation data. The paper concludes with a discussion of the results in the context of current practice.

2.3 What is Targeted Marketing Data?

In 2011, the New York Times published an article about a man who discovered his daughter was pregnant when Target mailed coupons relating to maternity clothing and nursery furniture to their home. How did Target know that the daughter was pregnant before others in her household? Target assigns customers a unique ID to track their purchasing behavior over time. The firm also appends TM data to each customer. The TM data include "demographic information like your age, whether you are married and have kids, which part of town you live in, how long it takes you to drive to the store, your estimated salary, whether you've moved recently, what credit cards you carry in your wallet, and what websites you visit" [7]. From the combination of purchasing behavior and demographic data, Target can reliably model customer behavior. And in this case, Target was able to predict pregnancy before a father could.

The New York Times article sheds light on how firms use TM data to customize marketing campaigns to potential customers. There are several large firms in the U.S. that compile TM databases for this particular purpose. One of these firms, Epsilon, maintains a database of 250 million individuals aged 18 years and older. Epsilon compiles information for each individual using public data (e.g., birth certificates, property records, change of address forms), credit card transaction data, credit reporting data, email or internet marketing data, and other sources. Information about an individual's interests can be gleaned from the credit card transactions and marketing emails. For example, pet owners will sign up for promotional emails from one or more pet stores and travel aficionados will purchase

multiple domestic and/or international air tickets over the course of a year. A description of the types of data that is typically collected and sold by TM firms can be found in Epsilon's Consumer Guide to Direct Marketing [9]. The information that is most applicable to travel demand modeling applications includes the following:

1. household demographics (e.g., number of adults and number of children in the household, family composition, household income),
2. individual demographics (e.g., age, gender, marital status, education, occupation),
3. housing and property data (e.g., owner/renter status, length of residence, dwelling type, home market value, property lot size, living area square footage, home sale date),
4. aggregated automotive data (e.g., average number of cars, trucks, recreation vehicles, and motorcycles in ZIP+4 area), and
5. lifestyle clustering (i.e., systems for classifying households nationwide).

TM databases, such as the one maintained by Epsilon, contain the majority of household and individual demographic fields that are used in travel demand forecasting models. It is important to point out that Epsilon does not currently have employment data. It could be possible to predict employment using a variety of strategies, some of which include tracking changes in the household income over time or observing work trips with other third-party data.

2.4 Advantages of Using TM Data

TM data provide several advantages over U.S. Census data. They are described below.

TM data are available at a more disaggregated level than most Census data The conventional approach to synthesize populations, which was developed by Beckman et al. [2], involves integrating aggregate data from one source with disaggregate data from another. The aggregate data are typically drawn from Decennial Census data, usually in the form

of one-, two-, or multi-way cross-tabulations. The disaggregate data, on the other hand, usually come from the ACS Public Use Microdata Sample (PUMS), which is the only publicly available untabulated Census data about individuals and households. PUMS data are reported at their most detailed level in Public Use Microdata Areas (PUMAs), which each contain about 100,000 residents. The PUMS files contain only about 5% of the housing units and 5% of the population in group quarters in the U.S. [27]. The method developed by Beckman and colleagues for synthesizing populations uses the disaggregate data as "seeds" to create individual records that are collectively consistent with the cross-tabulations provided by the aggregate data [11]. TM data provide disaggregate microdata like the ACS PUMS, but the data are available for a majority of the U.S. population with full addresses. Thus, the TM data can be used to generate more robust and location-specific synthetic populations.

TM data are regularly updated One of the key limitations of existing travel demand models is that they are based on episodic household travel surveys and Decennial Census data, which are usually updated once every ten years. The ACS data are available yearly, which provides a benefit over the Decennial Census and HHTS in terms of timeliness, but these data are released at least 9 months after collection. For example, the 2012 ACS 1-year estimates were not available until September 19, 2013 [26]. Multiyear estimates take even longer. In contrast, TM data are updated regularly. Monthly or quarterly updates of household income, residential moves, spending patterns, births and deaths, and lifestyle clusters could be particularly valuable in analyzing how economic or political shocks impact the transportation industry, or how a particular change in infrastructure affected an area.

TM data are considerably less expensive than traditional data As noted earlier, the cost of obtaining a completed travel survey for one household in Atlanta was approximately $200. In contrast, the cost to obtain TM data for one household is approximately five cents (although the actual cost will vary as a function of which variables and how many records are purchased). To put this in perspective, a 10% sample of households in Atlanta could

be purchased for approximately $25,000. Even if this purchase occurred annually, the non-discounted cost of using TM data over a 10-year period would be 12.5% of the cost of a decennial household travel survey. The TM data would also represent a 10% sample of the households in Atlanta, compared to the current household survey that represents 0.5% of the households. A portion of the cost difference could be used to fund other non-traditional data collection. For example, data that gives information about trip-making behavior could be purchased.

TM data are a national database TM data are similar to the Decennial Census in that, in theory, TM data contain information about all households in the U.S. The fact that the database is national presents a unique opportunity for comparison studies across different areas in the U.S. without having to control for differences in data collection methods.

TM data contain more information than Census data TM firms have access to behavioral information like where households shop and what TV channels they order. They develop proprietary algorithms using this vast amount of data to segment customers into different clusters. Numerous inputs are used to create these segmentation variables, including demographic data, financial data, survey data, transaction data, behavioral and attitudinal data, and trigger data. Trigger data refers to life events (such as the birth of a child) that might cause an individual to move into a different lifestyle cluster.

An example of one of the lifestyle segmentations maintained by Epsilon is called Niches 2.0, which contains 26 cleverly-named clusters ranging from the young and wealthy "Already Affluent" to the least prosperous "Zero Mobility." As an example, Epsilon describes the "Easy Street" cluster as follows: "The households in this Niche are typically older, white collar and educated. They have grown children, possibly still living with them. All of the households within this Niche own their homes and have lived at the same address for 7 years or more. On average, their homes are worth about $250,000. They are more likely than the general population to have a pool and to own a vacation home." These lifestyle clusters, as well as the individual behavioral variables used to classify them, provide an opportunity to

incorporate behavioral preferences and attitudes that are consistently available nationwide directly into travel demand models.

TM data can be used in longitudinal studies TM data are associated with individuals. Over time, many characteristics associated with individuals change: they marry, have children, get divorced, change residences, engage in new hobbies, etc. TM data can be used to examine the impacts of these, and other longitudinal changes, on travel behavior. Further, these types of longitudinal studies can be conducted at a national scale and at a substantially lower cost than traditional longitudinal studies. Issues associated with attrition are also expected to be less than those experienced with traditional longitudinal studies.

2.5 Disadvantages of Using TM Data

Before deciding to use TM data as a supplement to or an eventual replacement for HHTSs or certain types of U.S. Census data, there are several risk factors to consider.

Some TM fields are based on proprietary algorithms TM firms have propriety algorithms that they use to impute missing data or to create variables like lifestyle clusters. To remain competitive, many TM firms have developed proprietary algorithms to more accurately predict household income and lifestyle variables. Although the U.S. Census Bureau also imputes missing data with similar algorithms that are not readily available either, we cannot assume that a TM firms' imputation methods are as robust as those used by the U.S. Census Bureau. In our experience though, TM firms are continually seeking to improve their prediction models and are responsive to their clients' needs. For example, over the past two years, the TM firm we worked with made substantial improvements to the models it uses to predict the number of children in the household based on feedback from its clients (including us) that this was an important field.

Although the exact algorithms the TM firms use to populate these fields are not known, high-level details can often be shared with researchers. For example, the TM firm populates its household income variable with a model that is recalculated at least quarterly, and its

underlying algorithm is rebuilt every few years. Approximately 25% of the income data are obtained via credit applications and other self-reporting sources. Income for the remaining records is imputed using an algorithm that considers household data such as age, home ownership, home value, presence of children in the household, occupation, and education. In addition, it should be noted that TM firms likely use Census data as a part of their algorithms.

Without knowledge of the exact algorithms, researchers can test whether the results of the algorithms create a biased sample or whether updated algorithms result in different predictions. Changes in algorithms can produce incomparable data from one year to the next, but this type of problem is not specific to TM data. For example, since the 2010 Decennial Census the classifications of ethnicity have changed and therefore have made comparisons through time difficult. Nonetheless, researchers and practitioners should be aware that using TM data as a primary source for demographic data may require more frequent model calibrations, particularly when the TM firm updates its proprietary algorithms.

TM firms may go out of business or decide not to provide the data in the future Marketing is a well-established and thriving industry. In 2010, spending on advertising was estimated at $142.5 billion in the U.S. and $467 billion worldwide [8]. Multiple firms produce TM data, and actively compete to sell this data for use in direct marketing campaigns. Competition among TM firms bodes well for the travel demand modeling community, as it will keep TM data costs low. The presence of multiple TM firms, and their strong linkage to the marketing industry, also reduces the risk of relying on a data source that may be here today, but gone tomorrow. Although it may be tempting to estimate travel demand models that leverage the rich attitudinal and preference variables available through transactional purchase histories, researchers and practitioners need to assess the probability that these same fields will be available in the future. Developing models that rely on standard sociodemographic variables along with marketing variables such as the lifestyle segments (rather than highly-detailed variables) will reduce the risk that the data will not be available in the future.

In addition, third-party data providers may be impacted by future legislations that aim

to protect privacy. Although targeted marketing data has been sold for many decades, initiatives are underway that may make it possible for individuals to update and correct their information or to opt out [1, 10].

TM data are not perfectly representative of the U.S. population TM firms attempt to compile a database of all adults 18 years and older in the U.S. with a credit history. We expect that the TM database will not be completely representative of the U.S. population, and will underrepresent homeless, immigrants, and other special populations. *A priori*, we would also expect that the TM database would underrepresent individuals who have little to no credit histories, such as young adults and low income households. As with any dataset, selection bias should be expected. Researchers and practitioners should be aware that these biases may exist and should account for these biases in their models.

2.6 Assessing the Representativeness of TM Data

Before using TM data in travel demand models, it is important to assess whether the individuals and households contained in the TM data are representative of the population. In this section, we compare the distributions of 11 variables obtained from a January 2013 TM dataset to those reported in two sets of U.S. Census Bureau data: the 2010 Decennial Census and the 2007-2011 American Community Survey (ACS) 5-year estimates. Note that this was the most up-to-date ACS data available at the time of analysis. The 2008-2012 ACS 5-year estimates were not released until December 17, 2013 [26]. This study only compared data from one TM firm, but future research will extend the analysis to multiple TM firms.

The Decennial Census and ACS estimates each have associated measurements of error, which should be considered when interpreting this comparison. The Decennial Census' level of accuracy is estimated using the Census Coverage Measurement (CCM) survey, a post-enumeration survey that re-surveys a random sample of Census blocks throughout the country. The ACS's level of accuracy is reported within the tabulations themselves as margins of error. Future research will aim to quantify the effect of these inaccuracies as

related to TM data, but the scope of this study limits the comparisons to the estimates alone.

Given we expect to find some differences between the TM and Census data, in parallel we compare the ARC's 2011 HHTS to the same Census data. We compare the weighted and unweighted HHTS data. The weighted sample of ARC's HHTS incorporates two corrections into one weight: (1) an adjustment for the stratified sampling, and (2) a "raking" adjustment that aligns the sample to population statistics from 2008-2010 ACS 3-year estimates and the 2010 Decennial Census [19].

2.6.1 Variables

We purchased frequency distributions for 2,230 Census block groups for the 11 variables shown in Tables 1 and 2 from Epsilon. This represents all but two of the block groups in the 13-county metropolitan Atlanta region. One of the two block groups has a zero population according to the Census, and the TM firm inadvertently omitted the other.

TM and Census data use different categories for several of the individual and household variables. For the variables shown in Table 1, we were able to create categories that were directly comparable between the two datasets. For some variables, such as age, the TM data provided more refined categories whereas for other variables, such as income, the Census data provided more refined categories. The categories and associated mappings between the TM and Census data are shown in Table 1. For age, we created groups using multiples of 5 between 18 and 85 (i.e., 18-19, 20-24, 25-29, 30-34, ..., 85+). We used the block group as the unit of analysis when possible. However, because the Census data only provide the distribution of some variables at the tract level, we aggregated the TM data to tract level for these instances. These variables are noted in the right-hand column of Table 1.

There were two variables for which we could not create categories that could be directly compared. These variables, representing household size and length of residence, are shown in Table 2. We purchased TM data for household size as two separate tables: one for the number of adults in the household, and a second for the number of children in the household. Because the data was aggregated at the block group level, we could not meaningfully

Table 1: Variables that are Comparable between TM and Census Data

Variable	TM Categories	Census Categories	Census Table	Aggregation Level
Age	Exact ages 18-105	18-19 20, 21, 22-24, 25-29 30-34, 35-39 40-44, 45-49 50-54, 55-59 60-61, 62-64, 65-66, 67-69 70-74, 75-79 80-84, 85+	Table P12 2010 Decennial	Block Group
Education	Some HS or less High school Some college College Graduate school	No HS diploma HS degree, GED Some college, no degree Associates or Bachelors Graduate or professional	Table B15001 2007-11 ACS 5-yr	Tract
Ethnicity	African American Asian European Hispanic Other[a]	African American/Black Asian White Hispanic Other	Table P9 2010 Decennial	Block Group
Gender	Female Male	Female Male	Table P12 2010 Decennial	Block Group
Income	$0-15k $15k-20k $20k-30k $30k-40k $40k-50k $50k-75k $75k-100k $100k-125k $125k-150k $150k-175k, $175k-200k $200k-250k, $250k+	<$10k, $10k-15k $15k-20k $20k-25k, $25k-30k $30k-35k, $35k-40k $40k-45k, $45K-50k $50k-60k, $60k-75k $75k-100k $100k-125k $125k-150k $150k-200k $200k+	Table B19001 2007-11 ACS 5-yr	Tract
Marital Status	Married Single	Now married Never married, widow, divorced	Table B12002 2007-11 ACS 5-yr	Tract
Presence of Children	Yes No	Owner/renter w/ kids Owner/renter w/o kids	Table H19 2010 Decennial	Block Group
Tenure	Definite/probable owner Definite/probable renter	Owned w/,w/o mortgage Renter occupied	Table H4 2010 Decennial	Block Group

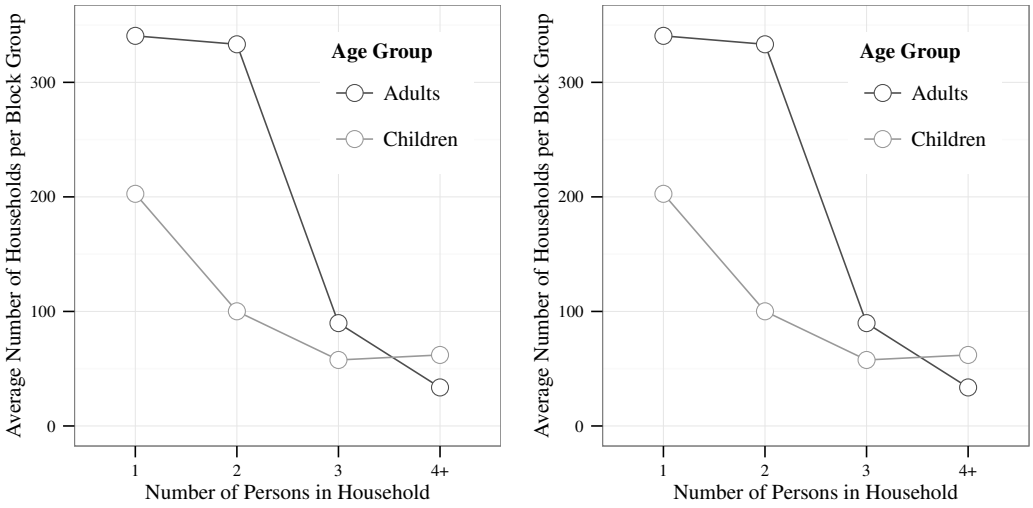
[a] Includes Jewish, Middle Eastern, Native American, and Oceanic.

Table 2: Variables that are Not Directly Comparable between TM and Census Data

Variable	TM Categories	Census Categories	Census Table	Aggregation Level
Num Adults	1,2,3,4,5+	n/a	n/a	n/a
Num Children	1,2,3,4+	n/a	n/a	n/a
HH Size	Available, not purchased	1,2,3,4,5,6,7+	Table H13 2010 Decennial	Block Group
Length of Residence	0-6 months	Moved in 2005 or later (0-6 yrs)	Table B25038 2007-11 ACS 5-yr	Tract
	7-12 months	Moved in 2000-2004 (7-11 yrs)		
	1-2 years	Moved in 1990-1999 (12-21 yrs)		
	3-5 years	Moved in 1980-1989 (22-31 yrs)		
	6-10 years	Moved in 1970-1979 (32-41 yrs)		
	11-15 years	Moved in 1969 or earlier (42+ yrs)		
	16-20 years			
	20+ years			

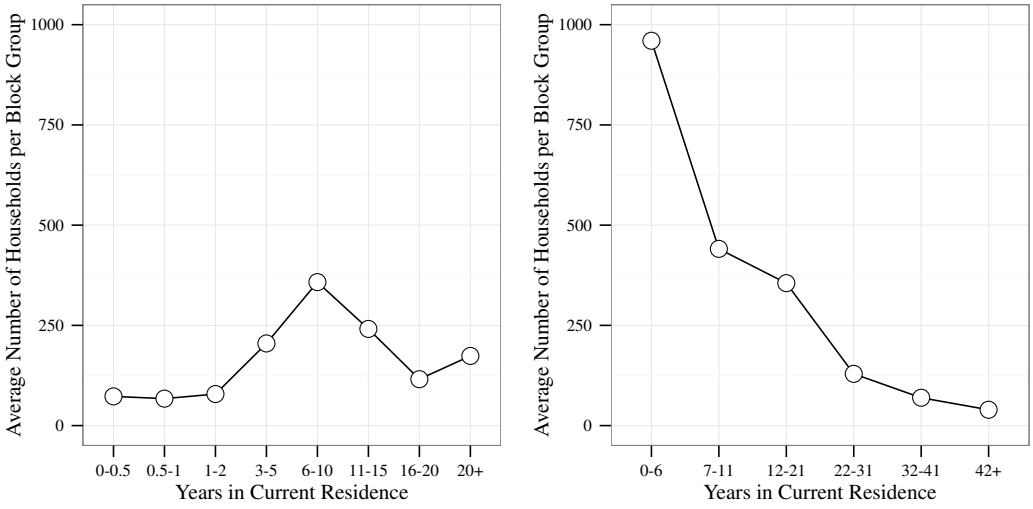
combine these two tables into a single one representing the total household size because after the aggregation we could no longer determine which households had children along with the adults. Furthermore, individual tables for the number of children and the number of adults are not provided as separate Census tables. For completeness, the distributions of the number of adults, number of children, and household size are shown in Figure 1. In future studies, this discrepancy can be avoided by purchasing the total household size from the TM data.

The other variable that is not directly comparable is length of residence. As shown in Table 2, the time scales used to represent the length of residence are quite different between the TM data and Census data. The TM data focuses on short-term movements, classifying any length of residence higher than 20 years into one group, whereas the ACS data provide more refined categories for residences beyond 20 years. Additionally, the cutoff years for the middle categories lag by one between the two data sets, which makes the comparison even more difficult. Based on a comparison of the distributions of the lengths of residences shown in Figure 2, we anticipate that the TM categories would be helpful for planning and modeling applications, as they provide researchers with the ability to understand short-term household movements.



(a) TM (b) Census

Figure 1: Household size.



(a) TM (b) Census

Figure 2: Length of residence.

2.6.2 Sample Size and Coverage Rate

We compared the data in both magnitude (meaning the overall sample size and coverage rate) and distribution (meaning the underrepresentation or overrepresentation of particular populations through coverage error). This section focuses on describing the methodology and results associated with the overall sample sizes and coverage rates. The coverage error associated with the frequency distribution of each variable is discussed in the next section.

2.6.2.1 Methodology

To understand how the total population of the TM data compares with Census estimates and the HHTS, the sample sizes and coverage rates are obtained for the study region. The sample size is defined as the raw number of records (either persons or households) in the data, and the coverage rate is defined as the sample size over the total estimate of persons or households by the Decennial Census, expressed as a percentage. These definitions conform to those of the U.S. Census Bureau [24, 25]. Additionally, the percent of complete records over the 11 variables is obtained. A "complete record" is defined as a person or household with non-missing data for every one of the 11 variables used in this analysis.

2.6.2.2 Results

Table 3 displays the overall sample size and coverage rates of each of the data sources. Note that all of the datasets were at least partially imputed at different points during data collection and processing so the percent of complete records is somewhat arbitrary. Missing HHTS data are imputed when they can be logically determined from other provided data [19]. Missing TM data are imputed for some variables using proprietary algorithms. Missing data in the ACS and Decennial Census are fully imputed [6, 23].

The coverage rate provided over the study area by the TM data is 87.5% at the individual level and 105.7% at the household level. A part of the household-level discrepancy could be due to the 3-year time difference between the TM data and Decennial Census data with new housing units being constructed during this time lapse. However, under this same argument the population grew as well, which would affect the individual-level coverage rate

Table 3: Total Sample Size and Coverage Rate of Data Sources As Compared to the Decennial Census for 13-County Atlanta, Georgia

Data Source	Sample Size	Coverage Rate	Complete Records	Year
		Persons		
TM	2,926,229	87.5%	—[a]	2013
HHTS	17,297	0.5%	95.9%	2011
HHTS weighted	16,576	0.5%	95.8%	2011
ACS (surveyed over 1 yr)	50,152[b]	1.5%[b]	94.2%[c]	2011
ACS 5-yr estimates	3,303,330	98.8%	—[d]	2007-11
Decennial Census	3,343,453	100.0%	87.3%[d,e]	2010
		Households		
TM	1,777,795	105.7%	92.6%[f]	2013
HHTS	8,971	0.5%	89.5%	2011
HHTS weighted	8,957	0.5%	89.7%	2011
ACS (surveyed over 1 yr)	25,224[b]	1.5%[b]	94.8%[c]	2011
ACS 5-yr estimates	1,639,172	97.5%	—[d]	2007-11
Decennial Census	1,681,614	100.0%	—[d]	2010

[a] Cannot be estimated from the data purchased for this study.
[b] In Georgia for 2011, 48,893 housing units responded to the ACS. According to the 2010 Decennial Census, there were 3,281,737 housing units in the state. Therefore, the coverage rate for the 13-county region is estimated as 1.5%. Sample size is estimated from the coverage rate.
[c] Estimated from national allocation rates [23].
[d] All missing data are imputed using logical or modeling imputation methods by the U.S. Census Bureau for final data products [6].
[e] Estimated from national data completeness statistic for person-level items [22].
[f] Estimated from a separate TM data purchase based on household-level data [14].

we calculated. It is more likely that the household-level discrepancy is due to the fact that the TM firm might not interpret individuals who indeed live together to be in the same household. Overall though, the coverage rate of the TM data is extensive, particularly if we consider it a sample whose underrepresented and overrepresented persons might be adjusted for with statistical weighting. Future research will examine the coverage rate by density or type of development such as urban, suburban, and rural.

2.6.3 Coverage Error

2.6.3.1 Methodology

A difference of percents measure is used to study the coverage error over variable categories. Coverage error, which includes both undercoverage and overcoverage, is defined as "the error in an estimate that results from (1) failure to include all units belonging to the defined population or failure to include specified units in the conduct of the survey (undercoverage), and (2) inclusion of some units erroneously either because of a defective frame or because of inclusion of unspecified units or inclusion of specified units more than once in the actual survey (overcoverage)" [12]. The difference of percents is used to compare the TM data to Census data and to compare ARC's HHTS data to Census data. The difference, d_i, is calculated for each category of each variable for every geographic area (either the block group or tract level, depending on the variable and data source). Note that the HHTS comparisons are all done at the tract level due to the geography of the traffic analysis zones used in reporting. We define d_i as:

$$d_i = \left(\frac{X_i}{\sum_{i=1}^{n} X_i} \right) 100 - \left(\frac{Y_i}{\sum_{i=1}^{n} Y_i} \right) 100 \tag{1}$$

where X is the frequency count (number of households or individuals) of category i in the TM or HHTS data, and Y is the equivalent for the Census data.

2.6.3.2 Results

The boxplots shown in Figures 3 and 4 depict six summary statistics (minimum, first quartile, median, mean, third quartile, and maximum) associated with the difference of percents, d_i, over geographic areas (block groups or tracts) for each variable category. The TM data, unweighted HHTS, and weighted HHTS as compared to Census data can be examined next to one another. The summary statistics are also listed in Tables 4 and 5.

Overall, the results show that the distributions of demographic and socioeconomic variables are similar between TM and Census data, particularly for age, gender, household

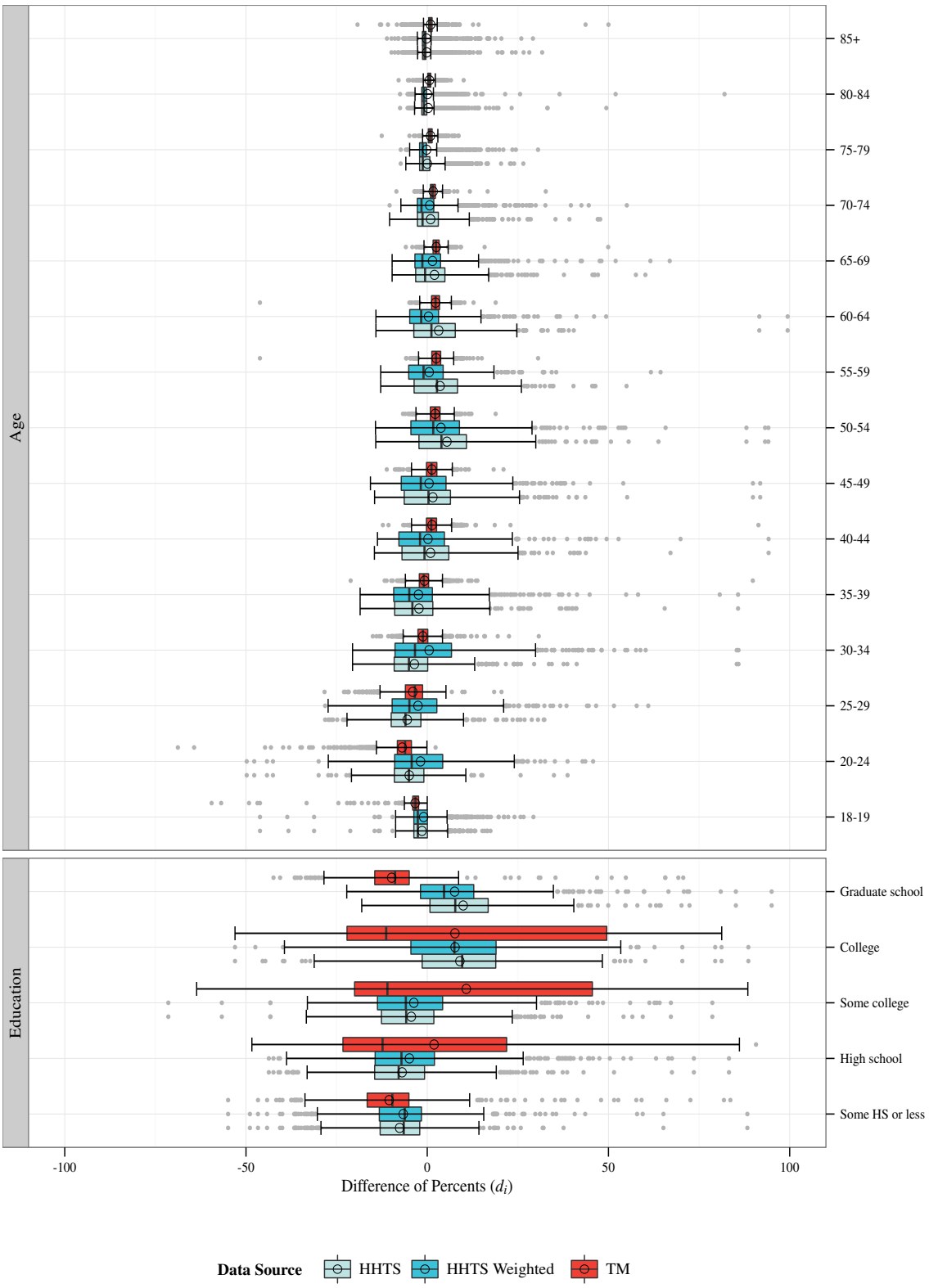


Figure 3: Boxplots displaying the distribution of d_i across data sources when compared to Census data.

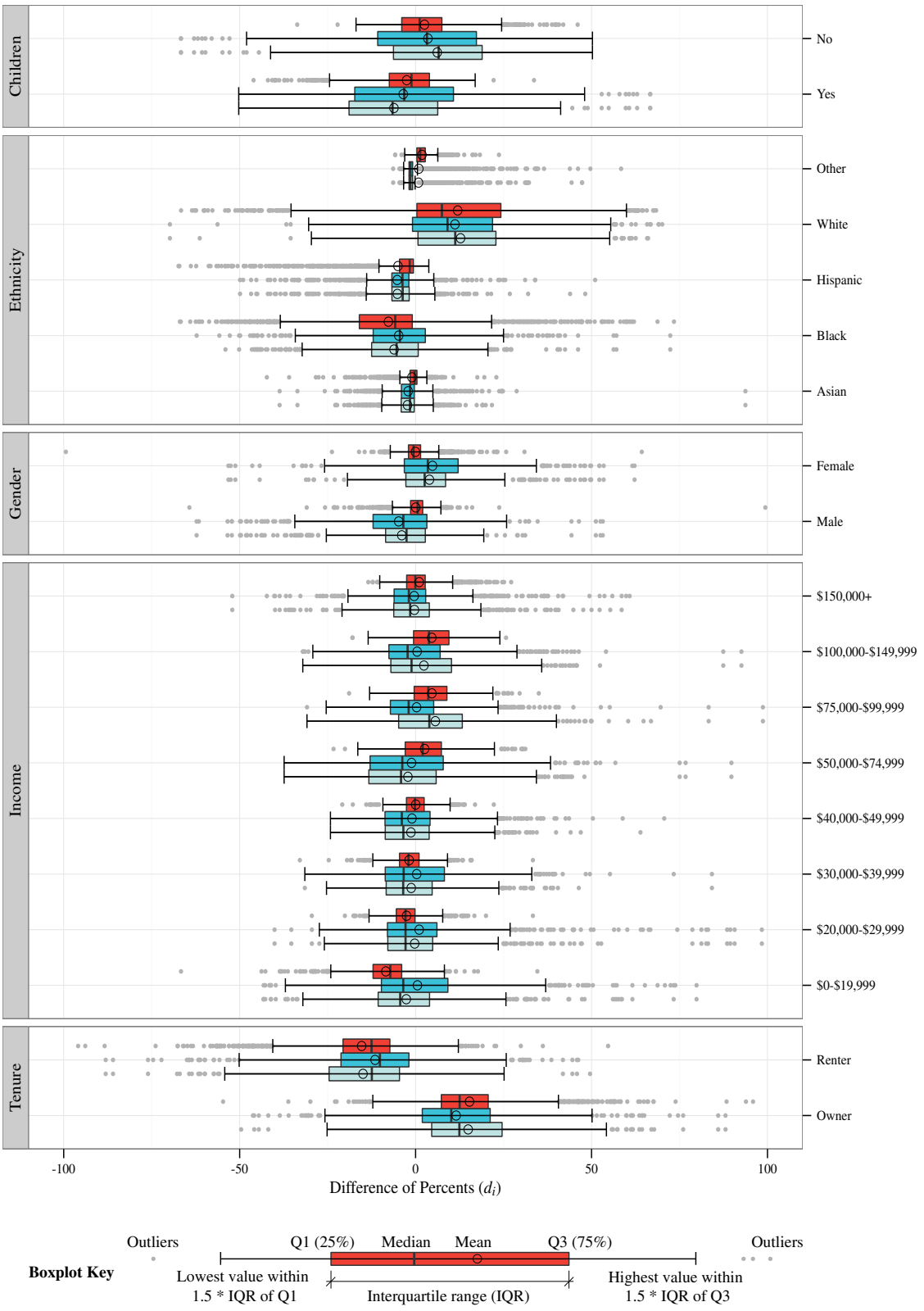


Figure 4: Boxplots displaying the distribution of d_i across data sources when compared to Census data.

income, and the presence of children. The largest discrepancies are associated with educational attainment and ethnicity. However, these discrepancies were comparable to those observed in the HHTS. The TM data contain fewer individuals under the age of 40, fewer low income households, and fewer households with children. The data are consistent with our expectations. Because the TM data are primarily derived from financial transactions, credit reporting data, and internet data, we expect those who have more access to credit or have longer credit histories to be better represented in the data. For all of the variables, however, the median differences observed at the block group or census tract levels are all within 12.5 percent, with the largest differences arising in housing type (tenure).

Due to the smaller sample sizes associated with the HHTS survey, the range of error of d_i is higher for the HHTS comparisons than the TM comparison in all cases except for educational attainment. Similar to the TM findings, the HHTS also contains fewer individuals under the age of 40, fewer low income households, and fewer households with children. Interestingly, the underrepresentation of renters and overrepresentation of Whites in both the TM and HHTS survey are almost identical. That is, the TM and HHTS survey both underrepresent renters and overrepresent Whites to a very similar degree.

The weighted HHTS data as compared to the unweighted HHTS data matches Census data somewhat more closely. Because the sample size of the HHTS was so small, the weighting has little effect on the overall distribution of d_i. However, if a similar weighting technique was adopted for the TM data, the biases present in the TM data have the potential to be reduced. Future research will investigate this hypothesis. Overall, we find that the distributions of demographic and socioeconomic variables that are commonly used in travel demand modeling applications are similar between the TM and HHTS survey.

2.7 Applications of TM Data

To illustrate how TM data can be combined with other data, we provide examples of research we have published or are working on that use TM data.

Travel demand sub-model In prior work, we used the TM data's lifestyle variables to predict

Table 4: Distribution of Difference of Percents, d_i, Across Data Sources

	TM						HHTS						HHTS Weighted					
Category	Min	25%	50%	Mean	75%	Max	Min	25%	50%	Mean	75%	Max	Min	25%	50%	Mean	75%	Max
AGE																		
18-19	-59.5	-4.0	-3.2	-3.3	-2.4	0.0	-46.1	-3.7	-2.6	-1.4	0.0	17.5	-46.1	-3.7	-2.6	-1.0	0.0	29.3
20-24	-68.8	-8.2	-6.1	-6.9	-4.4	2.3	-49.8	-9.1	-5.1	-5.0	-0.9	38.7	-49.8	-9.0	-4.3	-1.9	4.3	45.8
25-29	-28.3	-6.0	-3.4	-4.0	-1.3	20.5	-28.2	-9.9	-6.0	-5.5	-1.7	32.3	-28.2	-9.7	-4.9	-2.5	2.7	61.0
30-34	-15.0	-2.6	-1.2	-1.2	0.2	30.8	-20.6	-9.1	-5.1	-3.5	0.1	85.9	-20.6	-8.9	-3.4	0.5	6.7	85.9
35-39	-21.1	-2.2	-1.0	-0.7	0.4	89.8	-18.5	-9.0	-4.1	-2.3	1.5	85.8	-18.5	-9.2	-5.0	-2.4	1.4	85.8
40-44	-12.2	-0.2	1.1	1.3	2.6	91.4	-14.6	-7.0	-0.8	0.9	6.0	94.2	-13.7	-7.8	-2.0	0.2	4.7	94.2
45-49	-11.2	-0.2	1.2	1.3	2.7	21.1	-14.5	-6.3	0.3	1.6	6.4	91.9	-15.7	-7.2	-1.8	0.5	5.2	91.9
50-54	-6.7	0.9	2.2	2.2	3.5	18.9	-14.2	-2.3	3.9	5.4	10.8	94.1	-14.3	-4.5	1.7	3.8	8.9	94.1
55-59	-46.1	1.3	2.4	2.5	3.7	30.6	-12.8	-3.6	2.7	3.5	8.4	55.0	-12.8	-5.1	-1.0	0.5	4.4	64.4
60-64	-46.1	1.2	2.3	2.3	3.4	18.9	-14.2	-3.7	1.2	3.2	7.7	99.4	-14.2	-4.8	-1.7	0.4	3.1	99.4
65-69	-5.9	1.6	2.4	2.5	3.3	49.9	-9.7	-3.2	-0.6	2.0	4.9	60.2	-9.7	-3.4	-1.3	1.5	3.7	66.9
70-74	-8.5	0.9	1.5	1.6	2.2	32.7	-10.4	-2.7	-1.3	0.9	3.1	47.7	-10.4	-2.7	-1.6	0.7	1.8	55.1
75-79	-12.6	0.3	0.8	0.9	1.4	8.7	-7.3	-2.1	-1.2	0.0	0.7	26.5	-7.3	-2.1	-1.3	-0.2	-0.2	30.6
80-84	-7.9	0.1	0.5	0.6	1.0	10.1	-7.5	-1.5	-0.9	0.3	-0.1	49.4	-7.5	-1.5	-0.9	0.1	-0.2	82.0
85+	-19.2	0.4	0.8	0.9	1.3	50.0	-9.9	-1.3	-0.8	-0.1	-0.4	31.7	-11.1	-1.3	-0.8	-0.2	-0.4	29.2
EDUCATION																		
Some HS or less	-54.9	-16.6	-9.6	-10.5	-5.0	83.6	-54.9	-13.0	-6.5	-7.6	-2.1	88.3	-54.9	-13.2	-6.2	-6.6	-1.6	88.3
High school	-48.4	-23.2	-12.3	1.9	21.9	90.7	-43.7	-14.5	-7.9	-6.9	-0.7	83.2	-43.7	-14.4	-7.1	-4.9	2.0	83.2
Some college	-63.6	-20.1	-11.0	10.8	45.6	88.5	-71.4	-12.6	-5.8	-4.4	1.8	78.7	-71.4	-13.7	-5.9	-3.7	4.3	78.7
College	-53.0	-22.1	-11.3	7.6	49.5	81.2	-53.0	-1.4	9.6	9.0	18.9	88.6	-53.0	-4.5	7.5	7.7	19.0	88.6
Graduate school	-42.4	-14.5	-8.9	-9.8	-5.0	70.6	-18.1	0.8	7.7	9.9	16.8	95.0	-22.2	-1.9	4.6	7.6	12.8	95.0
ETHNICITY																		
Asian	-42.3	-1.6	-0.3	-1.0	0.4	22.9	-38.6	-4.1	-1.6	-2.3	-0.4	93.7	-38.6	-4.0	-1.6	-2.1	-0.4	93.7
Black	-67.0	-16.0	-5.8	-7.7	-0.9	73.4	-54.0	-12.5	-5.4	-6.1	0.7	72.4	-62.2	-12.1	-4.5	-4.8	2.8	72.4
Hispanic	-67.4	-4.6	-1.7	-5.0	-0.6	3.8	-49.9	-6.7	-3.7	-5.2	-1.8	48.2	-49.9	-6.7	-3.7	-5.2	-1.9	51.0
White	-66.7	0.4	7.5	12.0	24.2	68.4	-69.8	0.7	11.3	12.8	22.8	66.1	-69.8	-0.9	9.1	11.2	21.9	70.1
Other	-5.8	0.4	1.3	1.8	2.8	23.7	-6.4	-1.9	-1.4	0.8	-0.8	47.3	-6.4	-1.9	-1.4	0.9	-0.8	58.4

Table 5: Distribution of Difference of Percents, d_i, Across Data Sources

	TM						HHTS						HHTS Weighted					
Category	Min	25%	50%	Mean	75%	Max	Min	25%	50%	Mean	75%	Max	Min	25%	50%	Mean	75%	Max
GENDER																		
Male	-64.3	-1.4	0.6	0.0	2.0	99.4	-62.2	-8.5	-2.6	-3.9	2.8	53.2	-62.2	-12.1	-3.5	-4.8	3.2	53.2
Female	-99.4	-2.0	-0.6	-0.0	1.4	64.3	-53.2	-2.8	2.6	3.9	8.5	62.2	-53.2	-3.2	3.5	4.8	12.1	62.2
INCOME																		
$0-$19,999	-66.7	-12.1	-7.2	-8.5	-4.0	34.6	-43.2	-10.7	-4.4	-2.6	4.0	79.8	-43.2	-9.7	-3.5	0.5	9.2	79.8
$20,000-$29,999	-29.5	-5.5	-2.9	-2.6	-0.2	33.3	-40.1	-7.9	-2.9	-0.3	4.9	98.4	-40.1	-8.0	-2.9	1.0	6.1	98.4
$30,000-$39,999	-32.9	-4.6	-2.0	-1.8	1.0	33.3	-31.5	-8.4	-3.5	-1.2	4.7	84.2	-31.5	-8.6	-3.4	0.3	8.2	84.2
$40,000-$49,999	-20.8	-2.6	-0.2	0.0	2.4	22.2	-24.2	-8.6	-3.5	-1.3	3.9	63.9	-24.2	-8.7	-3.9	-1.0	4.1	70.7
$50,000-$74,999	-23.3	-3.0	2.0	2.6	7.3	31.2	-37.4	-13.3	-4.1	-2.2	5.8	89.8	-37.4	-12.9	-3.8	-1.1	7.9	89.8
$75,000-$99,999	-18.9	-0.4	3.6	4.6	8.9	35.0	-30.9	-4.8	3.9	5.6	13.3	98.7	-30.9	-7.2	-2.0	0.3	5.2	98.7
$100,000-$149,999	-18.0	-0.6	3.8	4.6	9.5	25.7	-32.1	-7.0	-1.2	2.4	10.2	92.6	-32.1	-7.6	-2.3	0.4	7.0	92.6
$150,000+	-13.4	-2.5	-0.1	1.0	2.7	27.2	-52.1	-6.2	-1.5	-0.3	3.9	58.6	-52.1	-6.1	-1.9	-0.4	2.9	60.8
PRESENCE OF CHILDREN																		
Yes	-46.1	-7.5	-1.2	-2.5	3.9	33.6	-50.3	-18.9	-6.5	-6.2	6.3	66.7	-50.3	-17.3	-3.2	-3.5	10.8	66.7
No	-33.6	-3.9	1.2	2.5	7.5	46.1	-66.7	-6.3	6.5	6.2	18.9	50.3	-66.7	-10.8	3.2	3.5	17.3	50.3
TENURE																		
Owner	-54.7	7.3	12.5	15.3	20.6	95.9	-49.6	4.6	12.4	14.9	24.6	88.1	-46.2	1.9	10.1	11.5	21.2	88.1
Renter	-95.9	-20.6	-12.5	-15.3	-7.3	54.7	-88.1	-24.6	-12.4	-14.9	-4.6	49.6	-88.1	-21.2	-10.1	-11.5	-1.9	46.2

the number of home-based airport trips to Hartsfield-Jackson Atlanta International Airport [13]. The model that used lifestyle clusters predicted the average number of air passenger trips better than the traditional models that used income to emulate ARC's existing airport passenger models. In this example, TM data was used in combination with airport survey data, rather than Census income data with the survey data.

Residential location choice We have also been using lifestyle variables to predict residential location choices. The lifestyle segments are working particularly well in this context, in part because they are able to capture individuals' preferences to live near others like themselves and away from those most unlike themselves [14]. In this application, TM data are being used in combination with Census data and other third-party data (namely, mobile phone data). We show that a HHTS is not needed to model residential location choice.

Emissions failure model In this study, TM data were linked at the household level to the Atlanta inspection and maintenance (I/M) emissions test database maintained by the Georgia Department of Motor Vehicles using Institutional Review Board (IRB) protocols for confidentiality. The linked database was used to investigate how household demographics and vehicle characteristics are associated with emissions failures [3].

Influence of built environment characteristics on vehicle ownership In this study, we explore the role of historical exposure to built environment characteristics on vehicle ownership. By using address histories, we examine how prior built environment characteristics associated with where people previously lived influence vehicle ownership [18].

Willingness to pay for proximity to public transit In this study, household demographics and home prices from TM data are used to compare different methodologies for incorporating spatial correlation; these models are used to estimate homeowners' willingness to pay for proximity to public transportation infrastructure. Household demographics, which are normally simulated from tabulated Census data, are available directly for each homeowner

in TM data [17].

Amending stated preference surveys In this study, researchers used the TM data to obtain names and addresses for individuals who lived within a specific area. These names and addresses were used to conduct a stated preference survey for a transit application. The survey provided one of the first opportunities to compare the accuracy of fields in the TM data at the household level (e.g., accuracy of names, addresses, gender, HH income, etc.). The study, which focused on low-income neighborhoods, can be extended in future research to a representative population. In future studies, supplementing stated preference or revealed preference data with TM data could provide additional information that is not traditionally available by surveying [16].

These examples highlight how TM data have been used to study travel behavior. These are just a few examples of many possibilities that illustrate the potential of using third-party data for travel demand modeling studies. In addition, all of these studies were able to strike a balance between two often conflicting objectives. The first is the need to collect detailed information about an individual's travel patterns and associate it with the individual's sociodemographic characteristics. The second is the need to protect the individual's confidentiality. In our experience, finding the balance is not an insurmountable challenge, but often requires some clever solutions. Early discussions with IRB representatives can help facilitate a quicker resolution to finding solutions that ensure individuals' identities are protected and are confidential.

2.8 Conclusions

Many researchers have been exploring ways to use non-traditional data sources to understand travel behavior. In this paper, we described how TM data can be a source of demographic and socioeconomic data. Our analysis suggests that TM data are similar to Census data, and is no more biased than the household travel surveys used today to build travel demand forecasting models. However, TM data have lifestyle and other behavioral

information that are not available in Census data or traditional HHTS. The inclusion of lifestyle and other behavioral information, combined with the ability to track individuals over time, provide the opportunity to examine many new research questions. Furthermore, for the great majority of MPOs that continue to maintain aggregate four-step travel demand models, which utilize simple medians or means, the TM data would perform as well as household travel survey data with a much larger sample size for sociodemographic information. For individual- or household-level data used with more detailed or advanced modeling, TM data could be weighted using traditional methods similar to HHTSs to correct for the underrepresented or overrepresented populations.

Looking ahead, the combination of targeted marketing data with other third-party and non-traditional data could be particularly powerful. Data from communication technologies offer the potential for researchers to better understand how instant information through the internet and our mobile phones influences (and can potentially be used to modify) travel behaviors. Movement and pattern data from mobile phone signaling and GPS providers offers the potential to provide real-time travel information to MPOs as well as accurate pictures of travel in the past. Combinations of these types of data offer tremendous opportunities to enhance, or even transform, existing travel demand modeling systems and data collection practices. Inexpensive, up-to-date, and detailed data available at regular intervals as often as every month or quarter would allow researchers to better study particular economic, climate, or political shocks in addition to transportation infrastructure changes. Such detailed information and sensitive modeling capabilities will be highly desirable in light of the new performance requirements in MAP-21 that will transform the federal surface transportation program to be more focused on performance outcomes.

2.9 Acknowledgements

Partial funding for this research was provided by a National Science Foundation Graduate Research Fellowship.

2.10 References

[1] Acxiom. "About the Data." https://aboutthedata.com, 2013. Accessed on 14 November 2013.

[2] R. J. Beckman, K. A. Baggerly, and M. D. McKay. "Creating Synthetic Baseline Populations." *Transportation Research Part A: Policy and Practice*, 30(6):415–429, Nov 1996.

[3] S. Binder, G. S. Macfarlane, L. A. Garrow, and M. Bierlaire. "Associations Among Household Characteristics, Vehicle Characteristics, and Emission Failures: An Application of Targeted Marketing Data." *Transportation Research Part A: Policy and Practice*, 59:122–133, 2014.

[4] California Department of Transportation. "2010-2012 California Household Travel Survey Final Report." http://www.dot.ca.gov/hq/tsip/otfa/tab/documents/chts_finalreport/FinalReport.pdf, 2013. Accessed on 14 November 2013.

[5] California Department of Transportation. "California Household Travel Survey." http://www.dot.ca.gov/hq/tsip/otfa/tab/chts_travelsurvey.html, 2013. Accessed on 22 July 2013.

[6] P. J. Cantwell, H. Hogan, and K. M. Styles. "The Use of Statistical Methods in the U.S. Census: Utah v. Evans." *The American Statistician*, 58:203–212, 2004.

[7] C. Duhigg. "The New York Times: How Companies Learn Your Secrets." http://nyti.ms/QbbTyS, 2012. Accessed on 22 July 2013.

[8] eMarketer Report. "Advertising Spending Statistics." http://www.statisticbrain.com/ad-spending-statistics/, 2012. Accessed on 24 July 2013.

[9] Epsilon. "Consumer Guide to Direct Marketing." http://www.epsilon.com/consumer-info/consumer-guide-direct-marketing, 2013. Accessed on 22 July 2013.

[10] Epsilon. "Consumer Preference Center." http://www.epsilon.com/consumer-preference-center, 2013. Accessed on 14 November 2013.

[11] J. Y. Guo and C. R. Bhat. "Population Synthesis for Microsimulating Travel Behavior." *Transportation Research Record: Journal of the Transportation Research Board*, 2014: 92–101, 2007.

[12] C. A. Konschnik. "Coverage Error in Establishment Surveys." In *Proceedings of the Survey Research Methods Section, American Statistical Association (1988)*, pages 309–314, 1988.

[13] J. D. Kressner and L. A. Garrow. "Lifestyle Segmentation Variables as Predictors of Home-Based Trips for Atlanta, Georgia Airport." *Transportation Research Record: Journal of the Transportation Research Board*, 2266:20–30, 2012.

[14] J. D. Kressner and L. A. Garrow. "Assessing the Viability of Lifestyle Clusters from Credit Reporting Data as an Alternative Dataset for Residential Location Choice Modeling." In *13th International Conference of the International Association for Travel Behavior Research (IATBR)*, 2012.

[15] J. D. Kressner and L. A. Garrow. "Using Third-Party Data for Travel Demand Modeling: A Comparison of Targeted Marketing, Census, and Household Travel Survey Data." *Transportation Research Record: Journal of the Transportation Research Board*, 2014. Accepted.

[16] J. D. Kressner, M. F. Carragher, and K. E. Watkins. "A Household-Level Pairwise Comparison of Targeted Marketing Data and Self-Reported Survey Data." In *Proceedings of the 2014 Annual Meeting of the Transportation Research Board*, 2014.

[17] G. Macfarlane, L. A. Garrow, and P. L. Mokhtarian. "The Influence of the Built Environment on Vehicle Ownership Preferences.". Georgia Institute of Technology. Working paper, 2013.

[18] G. Macfarlane, L. A. Garrow, and J. Moreno-Cruz. "Does Atlanta Value MARTA? Selecting an Autoregressive Model to Recover Willingness to Pay.". Georgia Institute of Technology. Working paper, 2013.

[19] PTV NuStats. "Regional Travel Survey: Final Report." Technical report, Atlanta Regional Commission, 2011.

[20] G. Rousseau. "Atlanta Regional Commission Transportation Coordinating Committee Meeting, 8 October 2010." 2010.

[21] State of California. "Strategic Growth Plan: Bond Accountability: California Household Travel Survey." `http://bondaccountability.resources.ca.gov/Project.aspx?ProjectPK=0540-OCA09017-4&pid=4`, 2013. Accessed on 22 July 2013.

[22] U.S. Census Bureau. "2010 Decennial Census: Item Nonresponse and Imputation Assessment Report." `http://www.census.gov/2010census/pdf/2010_Census_INR_Imputation_Assessment.pdf`, 2012. Accessed on 12 March 2014.

[23] U.S. Census Bureau. "American Community Survey Item Allocation Rates: Definitions." `http://www.census.gov/acs/www/methodology/item_allocation_rates_definitions/`, 2013. Accessed on 14 November 2013.

[24] U.S. Census Bureau. "American Community Survey Coverage Rates: Definitions." `http://www.census.gov/acs/www/methodology/coverage_rates_definitions/`, 2013. Accessed on 14 November 2013.

[25] U.S. Census Bureau. "American Community Survey Sample Size: Definitions." `http://www.census.gov/acs/www/methodology/sample_size_definitions/`, 2013. Accessed on 14 November 2013.

[26] U.S. Census Bureau. "American Community Survey 2012 Data Release." `https://www.census.gov/acs/www/data_documentation/2012_release/`, 2013. Accessed on 14 March 2014.

[27] U.S. Census Bureau. "2007-2011 PUMS Accuracy of the Data." `http://www.census.gov/acs/www/Downloads/data_documentation/pums/Accuracy/2007_2011AccuracyPUMS.pdf`, 2013. Accessed on 14 November 2013.

CHAPTER III

HOUSEHOLD-LEVEL VALIDATION FOR HARD-TO-REACH GROUPS

J. D. Kressner, M. F. Carragher, and K. E. Watkins. "A Household-Level Pairwise Comparison of Targeted Marketing Data and Self-Reported Survey Data." In *Proceedings of the 2014 Annual Meeting of the Transportation Research Board*, 2014

3.1 Abstract

This research conducts a validation test of targeted marketing data by comparing it at the household-level to self-reported survey data. The pairwise comparison was limited to the following demographic and socioeconomic variables: age, educational attainment, ethnicity, gender, household income, marital status, number of adults, number of children in the household, and tenure. The self-reported data were collected with a mailed stated preference (SP) survey regarding transit ridership in four neighborhoods of Atlanta that consist of many hard-to-reach and hidden populations. A rate of accuracy was calculated using a percent of correct matches between the two datasets for each variable. Chi-squared tests where also completed using both the targeted marketing and survey data. The findings suggest that targeted marketing data match self-reported data for neighborhoods of hard-to-reach or hidden populations at rates ranging from 17.4% to 94.5% depending on the variable. The self-reported data show that incorrect targeted marketing data randomly occur across all populations in relation to age, gender, household income, number of adults in the household, and tenure. It does not randomly occur across ethnicity or marital status groups. Educational attainment and the number of children in the household were not testable with regards to randomness across groups. Further research should be conducted to quantify the accuracy of targeted marketing data at the household-level for population groups that are more easily surveyed or documented.

3.2 Introduction

The survey-based data collection methods that became industry standard many years ago are still the standard today despite the rapid growth of new data types. Many advancement have been made regarding how surveys are collected, but even the most advanced household travel surveys conducted today still face declining response rates and smaller samples sizes each year. For example, the Atlanta Regional Commission 2011 Regional Travel Survey had a final response rate of 5.93% with a sample size of 0.5% [15]. Likewise, the California Department of Transportation (Caltrans) 2010-2012 Household Travel Survey had a final response rate of 2.0% with a sample size of 0.4% [1]. Additionally these surveys are collected infrequently, usually about every ten years. Urban population growth, the expansion of metropolitan areas, and the general unwillingness of the public to complete surveys conflict with our limited public funds, which has unfortunately resulted in a decline in the overall coverage of household travel surveys.

The Census Transportation Planning Package (CTPP) has also seen a decline in quality due to a combination of the following: (a) the relevant commute questions that are used to build Journey to Work matrices migrated from the late long form of the Decennial Census to the American Community Survey (ACS), and (b) the U.S. Census Disclosure Review Board (DRB) enacted minimum data dissemination requirements starting with CTPP 2000 [3]. Despite these facts, we still use the data from episodic household travel surveys and the CTPP for large portions of our travel demand models. Researchers and practitioners continually make improvements, particularly with the household travel survey [17], but ultimately budgets have limited these steps.

Even more, transportation planners have become attuned to the sensitivity of urban models, particularly with activity based models [5, 13, 14]. The validity and reliability of the demographic and socioeconomic inputs of these models are important [2], particularly when analyzing the model outputs disaggregately. And because the levels of demand placed by legislations on the abilities of travel demand models continue to rise, the detailed model outputs are scrutinized more and more closely. Many MPOs and other transportation planners or researchers find it difficult to pay for the current, detailed, and disaggregated

data that is required to build the kinds of urban models that are desired [7].

Meanwhile, we are living in a computer-driven world that is inundated with data. Third-party data are inexpensive, prolific, and information-rich. In a future scenario where household travel surveys no longer exist and the funding for the American Community Survey is cut [9, 16], third-party data could be our most promising source of up-to-date data. In particular, targeted marketing data could provide the demographic and socioeconomic data inputs for many travel demand and other urban modeling applications [10]. It could also append rich information to stated preference (SP) and revealed preference (RP) surveys.

Third-party data has largely been avoided in transportation modeling to date for a few reasons, the biggest of which seems to be the concern over the quality of the data. In order to address this concern, third-party data must be put through several validation tests that could identify inaccuracies or biases. This study offers one validation test, comparing targeted marketing data pairwise at the household-level to self-reported survey data for population groups that are typically undercounted, hidden, or hard-to-reach. The remainder of this paper is presented as follows: (1) a review of validation techniques used in transportation planning, (2) a description of the data used in this study, (3) a discussion of the methodology and results, and (4) concluding remarks with suggestions for future research.

3.3 Review of Validation Techniques

In 2007, the Transportation Research Board (TRB) Special Report 288 *Metropolitan Travel Forecasting: Current Practice and Future Direction* pointed out that model validation techniques are insufficiently emphasized and receive little effort when compared to other parts of the modeling process [6]. A subsequent report concluded that since input data drive urban models, quality control for the input data should be a part of model validation [2]. However, little direction is given on how to actually assess the validity of input data. Instead, it points out that currently the only consistent measure offered to decision-makers and the general public regarding model accuracy is in reference to the reliability of its travel forecasts in the base year rather than its ability to represent the population it aims to model. This

measure requires that the highway and transit model assignment results be compared with observed traffic volumes and transit boardings. Usually percent error calculations, defined in Equation 2, are reported in these comparisons.

$$\frac{|\text{Modeled Value} - \text{Measured Value}|}{\text{Measured Value}} \times 100\% \tag{2}$$

Equation 2 is not directly applicable to validation tests of input demographic and socioeconomic data because these data are generally categorical.

The *Online Travel Survey Manual* compiled by the TRB Travel Survey Methods Committee (ABJ40) has limited suggestions for validating survey data as well. It suggests that for surveys involving phone interviews, a small number of respondents should be recontacted to verify that they provide the same answers they did previously. For mail surveys, it is suggested that the survey ask respondents for telephone numbers so that they can be contacted again [18]. The manual also suggests validating the survey data with external sources such as U.S. Census Bureau data. However, it also specifies that U.S. Census Bureau data should be used to expand survey results to match the population using sampling weights and raking adjustments. By default this practice mirrors any trends present in the census data and therefore would render validation steps futile, providing a false sense of security. This discrepancy is acknowledged by MPOs, citing that validation steps are hampered by a dearth of independent data sources [6].

In this study, a unique opportunity exists that makes comparing categorical data between two completely independent data sources possible. The targeted marketing data are comparable house by house to respondents' self-reported data. Accordingly, a modified measure of accuracy is created that mimics Equation 2. This is discussed further in the methodology section.

3.4 Data

Two types of data are used in this study: targeted marketing data and self-reported survey data. The first dataset was obtained from a targeted marketing firm in April 2012. This dataset provided the sampling frame for the subsequent transit-related stated preference

(SP) survey, which was conducted from August to December 2012. The SP survey was conducted for two reasons: (1) to collect SP data on the perception and ridership impact of multi-modal transit mapping, and (2) to collect demographic and socioeconomic data for comparison against targeted marketing data. The former will be discussed in more detail in a forthcoming paper [4], whereas the latter is the focus of this paper. The researchers selected four neighborhoods within Atlanta to draw a sample from for the SP survey based on their proximity to bus routes that are included in the survey's multi-modal transit map. The neighborhoods are Pomona Park, Pine Lake, East Lake, and residents along Memorial Drive, which in total make up 11 U.S. Census block groups.

In general, the neighborhoods selected are not a representative sample of the Atlanta region. Rather, they primarily house hard-to-reach and hidden populations, such as those living in poverty, highly transient individuals or families, and undocumented persons. An analysis of data from the 2011 American Community Survey 5-year estimates showed that these areas have a lower mean annual household income ($43,336, with an average of 30.3% imputed records) than the city of Atlanta ($80,685, with 29.3% imputed) and the surrounding metropolitan area ($77,954, with 29.6% imputed). Additionally, 28.1% of the residents in the sampled areas are living in poverty. For comparison, 23.2% of Atlanta and 13.5% of the metro area live in poverty. The neighborhoods also have predominantly Black or African American populations (84.5% on average; city 54.1%, metro area 32.2%), a high percentage of zero car households (23.0% on average; city 17.8%, metro area 6.1%), and less people living in the same house that they did one year ago (72.4% on average; city 75.9%, metro area 82.3%).

3.4.1 Targeted Marketing Data

The researchers purchased data for 100% of the households in the 11 census block groups from a targeted marketing firm in April 2012. The dataset included 6,554 households. The percent of missing data for the variables used in this study are as follows: age 25.1%, educational attainment 0.3%, ethnicity 0.0%, gender 6.9%, household income 0.0%, marital status 0.0%, number of adults 0.0%, number of children 77.7%, and tenure 0.0%. The

number of missing records for the age and number of children variables are particularly higher than the other variables.

It should be noted that this particular targeted marketing data purchase cannot easily be compared to U.S. Census data because the Census geographies that the targeted marketing firm used as selection criteria were not yet updated to the 2010 boundaries at the time of the data purchase in 2012. In the four neighborhoods sampled, the census geography boundaries changed significantly between 2000 and 2010, and therefore any comparisons would need to be done with the 2000 data, making an unacceptable time difference of 12 years.

3.4.2 Self-Reported Survey Data

Out of the 6,554 households in the targeted marketing dataset, 2,000 households were randomly selected for the SP survey. The survey was conducted in three rounds. The first mailing was sent out in August 2012. This mailed letter included an online username and password asking that participants complete the survey online. In September, a reminder postcard was mailed out. Lastly, a paper version of the survey was mailed in November.

The demographic and socioeconomic questions included in the SP survey were formulated so that they are directly comparable to the data purchased from the targeted marketing firm. For example, the ethnicity categories provided in the SP survey matched those of the targeted marketing firm's categories exactly. Additionally, in the cases where too many categories were available in the targeted marketing data, they were aggregated so that the data was still comparable. For example, the income groups listed on the SP survey were "Less than $30,000," "$30,000-$49,999," "$50,000-$74,999," "$75,000-$99,999," and "$100,000 or higher," which shared break points with the income categories available in the targeted marketing data.

3.4.2.1 Completed Surveys

There were 116 participants that completed at least one of the demographic or socioeconomic questions, or 5.8% out of the original 2,000 mailed. For comparison, the 2011 ARC travel survey had a final response rate of 5.93%, and the 2010-2012 Caltrans survey had a

2.0% response rate as mentioned previously [1, 15]. The percent of missing data in the SP survey are as follows: age 12.1%, educational attainment 0.9%, ethnicity 3.4%, gender 0.0%, household income 5.2%, marital status 2.6%, number of adults 1.7%, number of children 0.9%, and tenure 1.7%.

Figure 5 shows the geographic distribution of the surveys that were completed versus those that were not completed. All of the mailings represented in this map were successfully delivered to the addressed household. Out of the 2,000 households in the sample, 64 of them were unable to be geocoded. These are absent from the maps. Only three of these 64 were completed surveys. There does not appear to be a significant geographic bias in survey respondents, who are shown in dark grey. However, there are quite a few instances where one building had a higher than expected rate of uncompleted surveys. These can be seen in the Pomona Park and Memorial Drive neighborhoods with the large blue circles that do not have correspondingly large dark grey circles.

The chi-squared tests using the targeted marketing data showed that there was a difference between those who filled out the survey and those who did not in income level only (p=0.0002). Households in the highest income group ($100,000+) responded to the survey more often than expected, and households in the lowest income group ($0-$29,999) responded less often than expected. The middle income groups responded as expected. The other variables' chi-squared tests produced the following non-significant results: age (p=0.1244, *simulated*), educational attainment (p=0.9549), ethnicity (p=0.6960), gender (p=1.0000), marital status (p=0.1830), number of adults (p=0.0548), number of children (p=0.4188, *simulated*), and tenure (p=0.1373).

3.4.2.2 Returned Mailings

It is important to point out that for 602 households out of the 2,000 in the sample (30.1%), at least one of the mailings out of the three rounds were returned to the researchers. The mailings were addressed to the name provided by the targeted marketing firm in hopes that having a specific name rather than "current resident" would improve the rate at which the mailing was opened. In doing so, the United States Postal Service did not deliver the

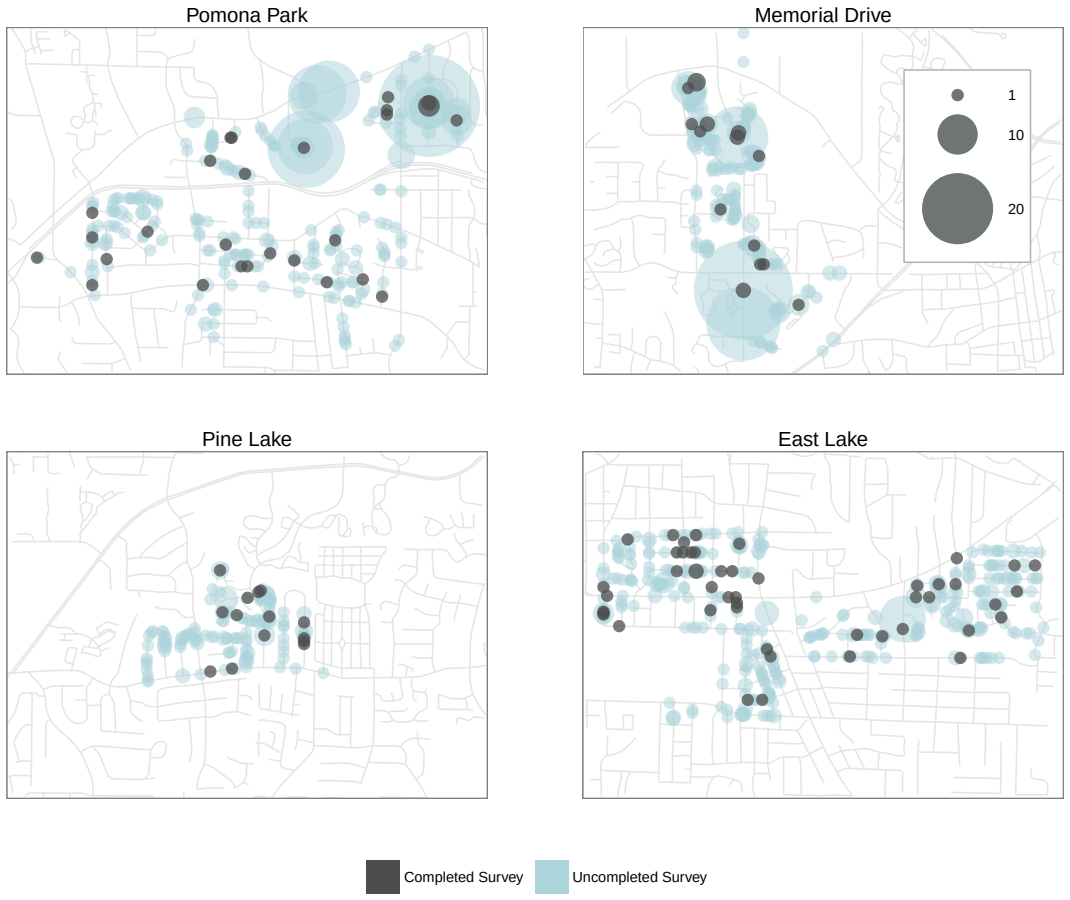


Figure 5: Locations of the completed versus uncompleted surveys for each neighborhood.

Table 6: Returned Mailings

Error Code		Frequency
Incorrect Name		
Not deliverable as addressed, unable to forward		246
Attempted, addressee not known, unable to forward		182
Moved, address known		108
Unable to forward or forward time expired		14
	Total	550
Incorrect Address or Vacant		
Vacant		27
No such street or number		13
Insufficient address		11
No mail receptacle		1
	Total	52

mailing if the addressee conflicted with information in their address and mail forwarding databases. In the sampled neighborhoods, an average of 27.6% of individuals moved within the past year alone according to the American Community Survey 5-year estimates (as stated previously), which makes it difficult for current address information to be maintained by targeted marketing firms. For primarily this reason, a high number of returned mailings occurred. Table 6 summarizes the mailing error codes. Note that 550 of these mailing error codes (91.4%) were valid addresses with other residents living in the household [19].

The researchers suggest that if future studies plan to use targeted marketing data, any mailings should be addressed to the name provided by the targeted marketing firm followed by "or current resident" in a subsequent line to avoid unnecessary returned mailings. If this approach is used, it would be advisable to ask at the beginning of a survey if someone currently lives in the household by the provided name and then, as a dependent question, if that person is the individual filling out the survey.

Figure 6 shows the locations of the returned mailings for each neighborhood. There are 14, of the 64 addresses that could not be geocoded (21.9%), missing from the maps. It is apparent from these maps that the number of returned mailings was very high for the Memorial Drive neighborhood. For East Lake, and possibly Pomona Park and Pine Lake, the returned mailings look geographically random when accounting for high density buildings. For example, in the upper right corner of the Pomona Park map, large dark grey circles (buildings with high rates of returned mailings) correspond with large orange circles

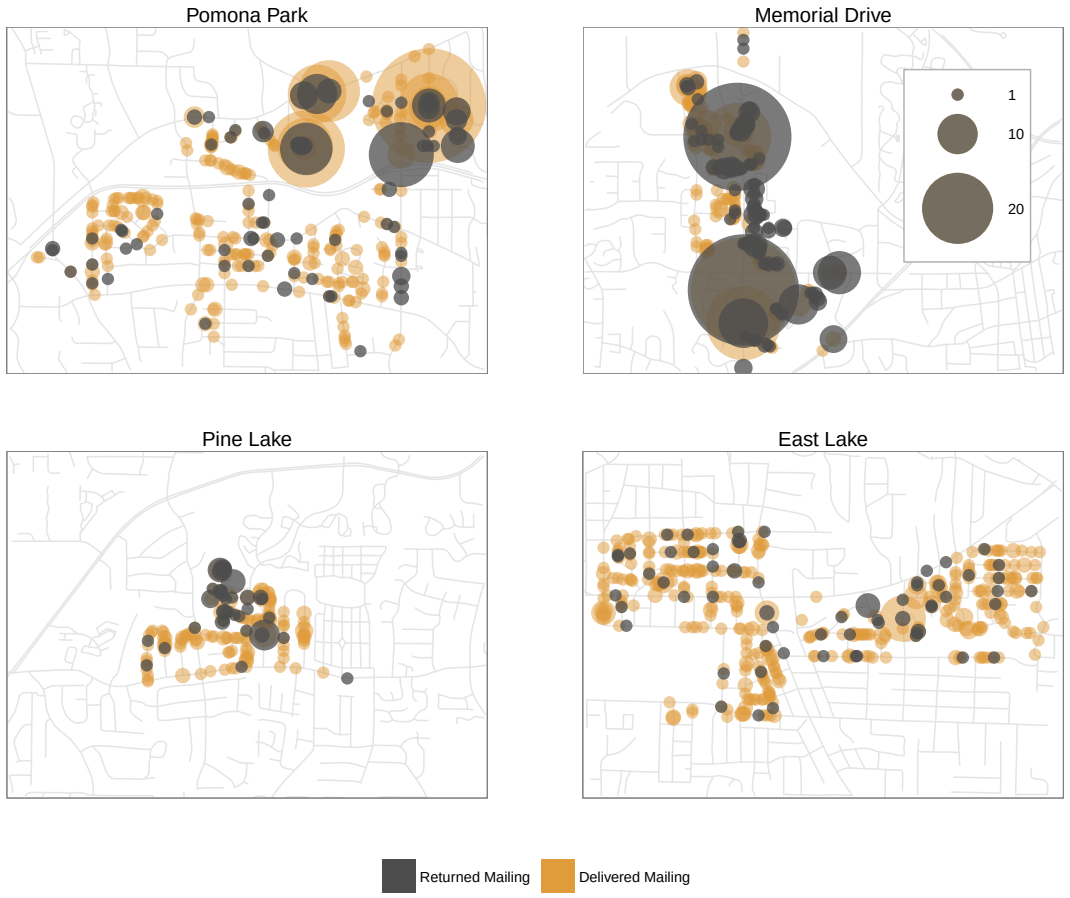


Figure 6: Locations of the returned versus delivered mailings for each neighborhood.

(buildings with high rates of successfully delivered mailings).

Chi-squared tests were run for each of the variables using the targeted marketing data to see if there was a difference between those whose mail was returned and those whose was successfully delivered. The results were significant for age (p=6.4e-11). There were more returned mailings than expected in the two lowest age groups (30 or under and 31-40). The opposite was true for the three highest age groups (61-70, 71-80, and 81 or over). This indicates that the targeted marketing firms have a harder time keeping a current address for younger households since younger individuals tend to move more often than older households. Correspondingly, the results for tenure were significant (p=2.2e-16), where a lot more mailings were returned for renters than expected. The reverse was true for owners. A related result could be seen in marital status (p=6.3e-15), where less mailings were returned

for married households than expected, and income (p=2.2e-16). There were more returned mailings than expected for the two lowest income ($0-$29,999 and $30,000-$49,999) and less than expected for the three higher income groups ($50,000-$74,999, $75,000-$99,999, and $100,000). Age, income, marital status, and tenure are all correlated.

Interestingly, the "traditional" households with two adults also received less returned mailings than expected. It is presumed that a majority of these households include a couple, whereas the one adult households, i.e. the single individuals, received more returned mailings than expected (p=0.0005, *simulated*). The results were also significant for ethnicity (p=0.0116) when comparing African American/Black, Caucasian/White, and all Other groups (these were combined due to small cell sizes). In this case, there were less returned mailings than expected for the Caucasian/White group.

The other variables' chi-squared tests produced the following non-significant results: gender (p=0.2717) and number of children (p=0.8373). The results were significant for educational attainment (p=2.2e-08), but only when testing the "High school completed" and "Some college" categories as the remaining were empty, which makes it unreliable. This is discussed in more detail in a following section.

3.5 Methodology and Results

3.5.1 Percent Correct

The self-reported data from the 116 completed SP surveys were compared pairwise to the corresponding households in the targeted marketing data at the household level. A simple `ifelse` statement is used to check the self-reported answer against the targeted marketing data for each variable house by house. If the data match, the record is labeled "Correct." Conversely, if they do not match, the record is labeled "Incorrect." If either or both of the households have missing data, the record is labeled "Missing." The results are summarized in pie charts in Figure 7 for each variable. The measure of accuracy that the researchers used, Equation 3, is shown in the middle of each pie chart.

$$\frac{\text{Correct Records}}{\text{Correct Records} + \text{Incorrect Records}} \times 100\% \tag{3}$$

Note that the missing records were removed from the calculation.

As an additional note, the researchers would like to point out that it is not possible to know with certainty whether the targeted marketing data was incorrect or whether the self-reported data was incorrectly answered, either intentionally or unintentionally. The assumption is made that the self-reported data is always correct. However, self-reported data are sometimes answered hastily and with privacy concerns at the forefront. Even more, for many households, some basic demographic and socioeconomic questions are difficult to answer within the parameters of a survey.

Gender, tenure, and age matched at 94.5%, 87.7%, and 82.3% respectively. Gender is likely the easiest of the variables tested to infer, so this makes sense. It is also relatively easy to understand the tenure of the household based on the type of housing that the current address is listed as. Furthermore, targeted marketing firms use data from credit reporting agencies that know definitively if an individual has or had a mortgage. Lastly, the age of individuals is usually obtained from credit reporting agencies, and so it again is relatively accurate. Note that a portion of the incorrect matches could be due to the fact that the "head of household" was subjectively defined by both respondents and the targeted marketing firm. Ages were obtained for up to five individuals in each household in the targeted marketing data, but a comparison was only done against Person 1. The respondents' opinion of who the head of household was could have been one of the individuals listed as Person 2-5.

Marital status and ethnicity matched at 69.0% and 63.4% respectively, which is fair, but the rate of matching for the remaining variables is seemingly too low (educational attainment, household income, number of adults, and number of children in the household). When examined more closely though, even the low accuracy of these variables does not render them useless. Because of the low cost of purchasing targeted marketing data, there is a lot of data. The number of households with correct data, which can be estimated from the results in Figure 7, is still orders of magnitude greater than the number of households from which survey data was collected through the extended survey effort. This fact is true for even the surveyed neighborhoods in this study with many hard-to-reach and hidden

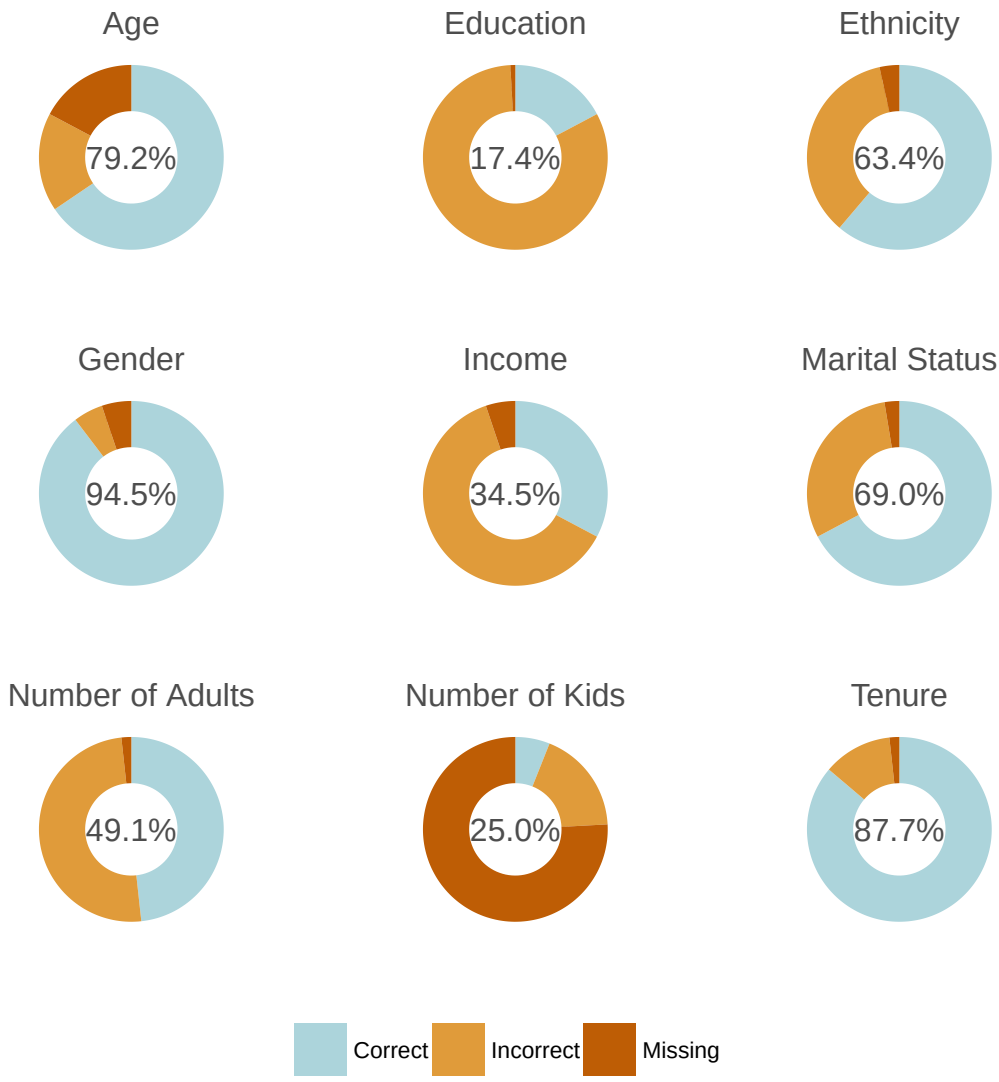


Figure 7: Results of the comparison between the targeted marketing and self-reported survey data. The percentage is calculated as the number of correct matches over the total number of records compared (excluding the missing records) for each variable.

populations.

To illustrate this point clearly, the household income variable will be used as an example. Excluding the missing records, 34.5% of the targeted marketing data's income records would be correct. That means that for the four neighborhoods in this study, 34.5% of the 6554 records are correct, or 2264 households in these neighborhoods. Compared to the 110 income records received from the SP survey, this number is much higher (1958% increase). The percent increase in sample size for the remaining variables is as follows: age (3530%), educational attainment (888%), ethnicity (3609%), gender (4872%), marital status (3903%), number of adults in the household (2724%), number of children in the household (217%), and tenure (4942%).

For a lower cost in both money and time, it is estimated that accurate data was purchased for a much larger sample of the population than could likely be attained with a survey, whether online, mailed, or door-to-door. If it can be determined for which households the data are correct, incorrect, or missing, the biases could be statistically corrected with weighting and imputation techniques. The problem, albeit a difficult one, is then to figure out for which portion of the population in each variable the data are correct and for which it is not. Is it randomly correct over the entire population, or is it more correct for certain income, age, or ethnicity groups than others?

3.5.2 Chi-squared Tests of Independence

To answer the question of randomness in incorrect data, the researchers examine the self-reported data more closely. Chi-squared tests of independence are used to find any existing relationships between the variables and data accuracy. In the cases where the population in each cell is too small (about <5), the test's p-value is computed using a Monte Carlo significance test procedure with $2,000$ replicates [8]. Both educational attainment and the number of children in the household were not tested. The educational attainment variable was imputed in such a way as to make many of the cells empty. Also, the variable regarding number of children in the household is largely missing, which makes an inquiry into its relationship with data accuracy futile.

Table 7: Chi-Squared Test of Independence for Income

	Correct	Incorrect
\$0-\$29,999	14	21
\$30,000-\$49,999	6	19
\$50,000-\$74,999	6	15
\$75,000-\$99,999	3	10
\$100,000+	9	7
p=0.1989, *simulated*		

Age (p=0.1609, *simulated*), gender (p=1.0000, *simulated*), income (p=0.2004, *simulated*), number of adults (p=0.4769, *simulated*), and tenure (p=0.2104, *simulated*) each fails to reject the null hypothesis, which assumes there is no relationship between the variable and data accuracy. It is particularly interesting that the data did not show a relationship between income groups and whether the record matched between the two data sets. Because household income is an important variable used in transportation modeling, understanding this variable's behavior in the targeted marketing data is imperative. It was expected that there would be an association because of the low rate of matching responses combined with the characteristically poorer neighborhoods surveyed. The failure to discover a relationship may be due to the small sample size of the survey. Table 7 shows the observed data and the p-value using Monte Carlo simulations. Although the null hypothesis could not be rejected, it appears that the only two income groups that had more incorrect matches than correct matches were the lowest and highest groups. With a larger sample, this tendency may prove to be statistically significant.

Both ethnicity (p=0.0005) and marital status (p=0.0000) did in fact reject the null hypothesis, asserting that there is a relationship or association with data accuracy. There are more households than expected whose targeted marketing and self-reported data matched for the African American/Black category. The reverse is true for the Caucasian/White category. The researchers hypothesize that the missing ethnicity data are at least partially modeled by the targeted marketing firm with Census data, and therefore it was estimated that most individuals living in these neighborhoods are African American/Black. This could have resulted in an association between ethnicity and data accuracy. On the other hand,

the association between marital status and the rate of matching responses is unlikely. There is a higher than expected occurrence of single individuals whose data matched and married individuals whose data did not. This could potentially indicate that the targeted marketing data assumes one is single until data are collected otherwise. According to this logic, the married individuals whose data did not match could have been married more recently.

3.6 Conclusions and Future Research

This study concludes that the rate of accuracy between the targeted marketing and self-reported data for neighborhoods with hard-to-reach and hidden populations is relatively high for age, gender, and tenure (ranging from 82.3% to 94.5%), but for educational attainment, ethnicity, household income, marital status, number of adults, and number of children in the household, it ranges from 17.4% to 69.0% (see Figure 7). Despite this, the estimated number of correct records obtained in the targeted marketing data for the four neighborhoods still surpasses the number of records obtained from the SP survey by a large amount. The self-reported data also show that incorrect targeted marketing data randomly occur across all populations in relation to age, gender, household income, number of adults in the household, and tenure. It does not randomly occur across ethnicity or marital status groups. Educational attainment and the number of children in the household were not testable with regards to randomness across groups. Future research should further test for which groups targeted marketing data are inaccurate, and it could additionally test the effectiveness of methods for predicting inaccuracies.

The present study examined the survey responses of a small sample of individuals (116 out of 6554 households, 1.8%) from just four targeted neighborhoods that included high rates of hard-to-reach and hidden populations (11 Census block groups out of 2232 for the 13 county planning region, 0.5%). The researchers seized the opportunity to look at this particular group of people, assuming it would be the worse case scenario regarding accuracy in the targeted marketing data. A separate study by the authors compares tabulated targeted marketing data with current Census data and a recent household travel survey for the entire 13-county region [11]. In general, the results show that targeted marketing

data closely represent Census data when comparing basic demographic and socioeconomic variables at an aggregate level. The median differences observed at the block group or tract levels are all within 12.5 percent, with the largest differences arising in tenure. From these results, we can hypothesize that targeted marketing data are more accurate for those not classified as hard-to-reach or undercounted groups. Future research should investigate the improvement of accuracy for these population groups.

3.7 Acknowledgements

This research is based upon work supported by the National Science Foundation Graduate Research Fellowship.

3.8 References

[1] California Department of Transportation. "2010-2012 California Household Travel Survey Final Report." `http://www.dot.ca.gov/hq/tsip/otfa/tab/documents/chts_finalreport/FinalReport.pdf`, 2013. Accessed on 14 November 2013.

[2] Cambridge Systematics, Inc. "Travel Model Validation Practices Peer Exchange White Paper." Technical report, Federal Highway Administration, 2008.

[3] Cambridge Systematics, Inc. "NCHRP 08-36 Task 71: Disclosure Avoidance Techniques to Improve ACS Data Availability for Transportation Planners." Technical report, American Association of State Highway and Transportation Officials, 2009.

[4] M. Carragher, K. E. Watkins, and J. D. Kressner. "Transit Rider Response to Multi-Modal Mapping.". Georgia Institute of Technology. Working paper, 2013.

[5] J. Castiglione, J. Freedman, and M. Bradley. "Systematic Investigation of Variability due to Random Simulation Error in an Activity-Based Microsimulation Forecasting Model." *Transportation Research Record: Journal of the Transportation Research Board*, 1831:76–88, 2003.

[6] Committee for Determination of the State of the Practice in Metropolitan Area Travel Forecasting. "Special Report 288 Metropolitan Travel Forecasting: Current Practice and Future Direction." `http://onlinepubs.trb.org/onlinepubs/sr/sr288.pdf`, 2007. Accessed on 25 July 2013.

[7] Federal Highway Administration and Federal Transit Administration. "Using ACS Data in Transportation Planning Applications." `http://planning.dot.gov/Peer/Daytona/daytona_2007.asp`, 2007. Accessed on 25 July 2013.

[8] A. C. A. Hope. "A Simplified Monte Carlo Significance Test Procedure." *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(3):582–598, 1968.

[9] House of Representatives, 113th Congress, 1st Session. "To make participation in the American Community Survey voluntary, except with respect to certain basic questions, and for other purposes, H.R. 1078." http://thomas.loc.gov/cgi-bin/query/z?c113:H.R.1078:, 2013. Accessed on 26 July 2013.

[10] J. D. Kressner and L. A. Garrow. "Lifestyle Segmentation Variables as Predictors of Home-Based Trips for Atlanta, Georgia Airport." *Transportation Research Record: Journal of the Transportation Research Board*, 2266:20–30, 2012.

[11] J. D. Kressner and L. A. Garrow. "Using Third-Party Data for Travel Demand Modeling: A Comparison of Targeted Marketing, Census, and Household Travel Survey Data." *Transportation Research Record: Journal of the Transportation Research Board*, 2014. Accepted.

[12] J. D. Kressner, M. F. Carragher, and K. E. Watkins. "A Household-Level Pairwise Comparison of Targeted Marketing Data and Self-Reported Survey Data." In *Proceedings of the 2014 Annual Meeting of the Transportation Research Board*, 2014.

[13] J. D. Lemp, L. B. McWethy, and K. M. Kockelman. "From Aggregate Methods to Microsimulation: Assessing Benefits of Microscopic Activity-Based Models of Travel Demand." *Transportation Research Record: Journal of the Transportation Research Board*, 1994:80–88, 2007.

[14] R. M. Pendyala and C. R. Bhat. "Validation and Assessment of Activity-Based Travel Demand Modeling Systems." In *Innovations in Travel Demand Modeling Conference, Transportation Research Board*, 2006.

[15] PTV NuStats. "Regional Travel Survey: Final Report." Technical report, Atlanta Regional Commission, 2011.

[16] Senate, 113th Congress, 1st Session. "To make participation in the American Community Survey voluntary, except with respect to certain basic questions, and for other purposes, S. 530." http://thomas.loc.gov/cgi-bin/query/z?c113:S.530:, 2013. Accessed on 26 July 2013.

[17] P. R. Stopher and S. P. Greaves. "Household travel surveys: Where are we going?." *Transportation Research Part A: Policy and Practice*, 41:367–381, 2007.

[18] Travel Survey Methods Committee (ABJ40). "Household Survey Data Expansion and Analysis." http://www.travelsurveymanual.org/Chapter-12-1.html, 2013. Accessed on 25 July 2013.

[19] United States Postal Service. "Domestic Mail Manual." http://pe.usps.com/archive/html/dmmarchive1209/F010.htm, 2013. Accessed on 20 July 2013.

CHAPTER IV

AIRPORT PASSENGER MODEL

J. D. Kressner and L. A. Garrow. "Lifestyle Segmentation Variables as Predictors of Home-Based Trips for Atlanta, Georgia Airport." *Transportation Research Record: Journal of the Transportation Research Board*, 2266:20–30, 2012

4.1 Abstract

This research investigated the influence of demographic and socioeconomic factors on air travel demand by using a unique data set purchased from a credit reporting agency. Linear regression models based on lifestyle segmentation variables were used to predict air passenger trips for Hartsfield-Jackson International Airport in Atlanta, Georgia. The study focused on predicting trips that originated from or terminated at residences in Atlanta's 13-county metropolitan area. The lifestyle regression models were compared with regression models based on income, because the latter were similar to the regression models currently used by the Atlanta Regional Commission to predict home-based airport passenger trips. The results provide directional evidence for using lifestyle clusters over income groups in predicting airport passenger trips. The evidence suggests that alternative data sources with adequate information for lifestyle segmentation can improve airport passenger models. The discussion points out the need for air passenger surveys to collect information about the number of annual air trips a surveyed individual takes.

4.2 Introduction

Two general types of model users and developers are interested in the demographics and socioeconomics of airport travelers: airport planners and metropolitan planning organizations. Whereas airport planners are generally interested in understanding passenger characteristics for planning and forecasting needs of the airport infrastructure and capacity, metropolitan planning organizations are primarily interested in air passenger characteristics to allocate

geographically and plan for the trips to their region's airports.

The Atlanta Regional Commission (ARC), which is the metropolitan planning organization for Atlanta, Georgia, maintains the region's travel demand model. This model is unique in that it estimates trips for Hartsfield-Jackson Atlanta International Airport in an airport passenger model [23]. Airport trips are not normally estimated in metropolitan travel demand models because the frequency of air travel from a household is so small that most home travel surveys, which are the primary data source in travel demand models, do not observe more than two or three air passenger trips. However, in many metropolitan areas, airports can generate a large proportion of trips during certain times of the day and days of the week.

To develop an airport passenger model that captures the large volume of trips being generated by Hartsfield-Jackson, ARC used databases available from the airport planning department and government agencies. These databases are commonly used for other aviation forecasting models. However, as noted in an ACRP problem statement, forecasting models based on these databases are limited because they use "very aggregate measures of demographic or economic factors, such as total population, gross domestic product ... or per-capita disposable income ... as the principal, and often only, independent demographic or socioeconomic variables" [4]. Motivated by these data needs, this research explores an alternative data source available through a credit-reporting agency that can provide disaggregate demographic and socioeconomic information about air passengers, including a lifestyle segmentation system.

4.3 Literature Review

4.3.1 Demographics in Air Travel

4.3.1.1 General Models

Four major forecasting methods are considered in airport activity forecasting: market share forecasting, econometric (regression) modeling, time series modeling, and simulation modeling [22]. Each of these forecasting methods can potentially use demographic and socioeconomic information to account for passenger characteristics in a variety of modeling

applications (e.g., choice models, no-show-rate models, and food sales predictions in the airport). Airport planning groups and metropolitan planning organizations conduct air passenger surveys for many of these forecasting needs, sometimes annually, but often detailed demographic and socioeconomic information is not collected with these surveys and the models suffer accordingly.

4.3.1.2 Atlanta's Airport Passenger Model

The ARC airport passenger model accomplished two main tasks: the distribution and mode choice of the air passenger trips to metropolitan Atlanta. ARC estimated the number of daily air trips and geographically allocated them using three data sources: the total number of enplanements at the airport as reported by the Federal Aviation Administration (FAA), an air passenger survey conducted at Hartsfield-Jackson Atlanta International Airport in 2000, and the 2000 decennial census [23].

Figure 8 summarizes ARC's air passenger distribution model. Several characteristics of the model are relevant in the context of this study. First, the number of airport trips was broken down by resident status, type of air trip, and non-airport trip end categories. Second, non-airport trip ends from the air passenger survey were allocated to traffic analysis zone (TAZ) linear regression models. The regression models associated with the home-based trip ends were based on the total number of households in each income group as reported in the 2000 decennial census. These regression models were constrained to ensure that trip rates increased with income. The regression models associated with the "other" trip end are based on total employment in the TAZ. ARC notes that the distribution portion of the air passenger model shown in Figure 8 did not offer a high degree of prediction accuracy, even after applying geographic indicators with related factors to the central business district, outlying counties, and an "other" category.

4.3.2 Life Cycle and Lifestyle Segmentation Literature

The terms "life cycle" and "lifestyle" appear in the literature frequently. Although these terms are sometimes used interchangeably, they refer to different concepts. In scientific studies involving a life process from birth to death, the term life cycle is used, where each

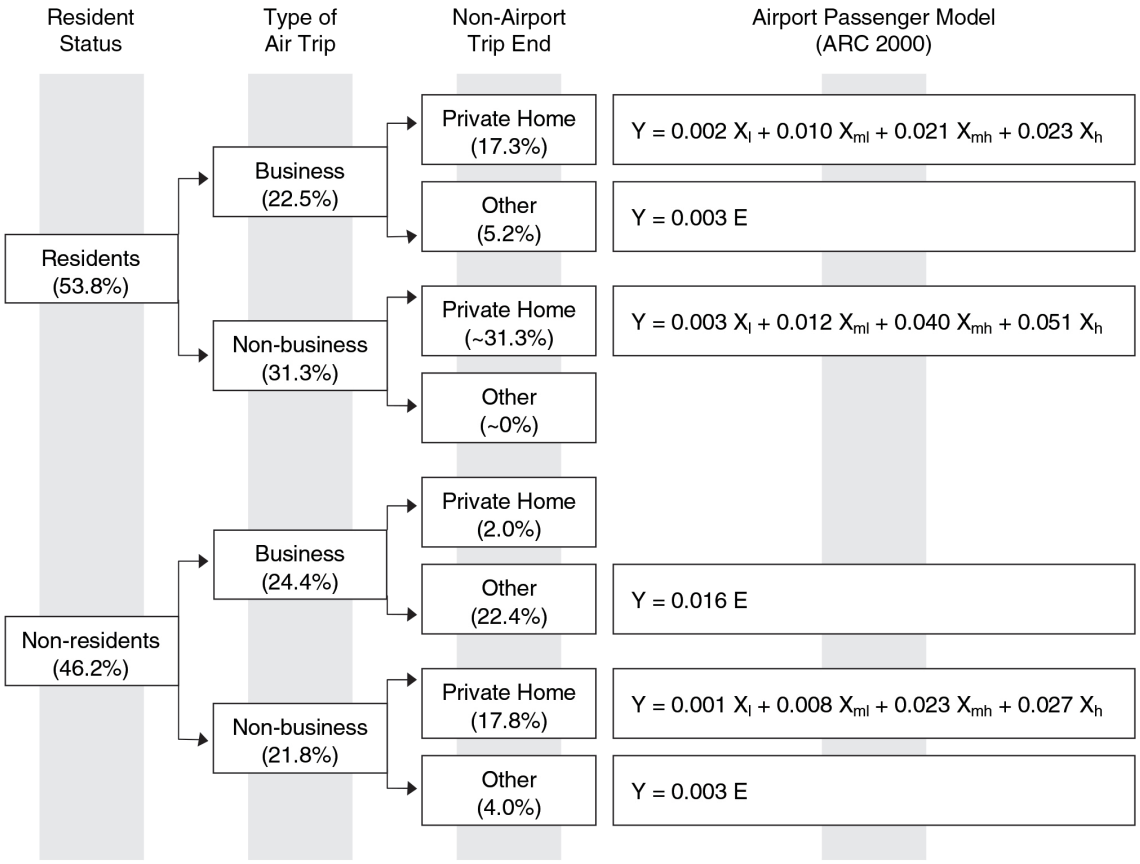


Figure 8: ARC airport passenger model.

Y = total daily air passengers per traffic analysis zone.
X = number of households in the income group as 1999 dollars; where X_l = low ($0 to $19,999), X_ml = medium low ($20,000 to $49,999), X_mh = medium high ($50,000 to $99,999), X_h = high ($100,000 and over).
E = total employment.

step in the life cycle is called a life stage. In general, models that account for demographic characteristics and socioeconomic factors are characterizing the life cycle. In contrast, models that incorporate taste variations beyond simple demographic and socioeconomic factors—for example, by including attitudinal, behavioral, and habitual information—are characterizing the lifestyle. A detailed discussion of these terms and an extensive literature review as well as earlier research are available elsewhere [9, 18, 19].

In practice, the distinction between life cycles and lifestyles is not easily decipherable. For example, one study created a "life cycle" segmentation scheme using variables that distinguished young and old households, the presence of children, and blue-collar versus white-collar occupations [2]. Another study created a "lifestyle" segmentation scheme using similar variables to describe the household structure (age of the head of household, presence of children and their age groups, number of adults, and household size), labor force participation (household income, employment status of the male and female heads of household, proportion of white-collar male employees, and variables explaining the proportion of income made by both the male and female), and leisure activities [20]. The leisure activities, as described by the authors, included variables such as education levels rather than actual behavioral data. The difference between the two studies, therefore, rests more on the level of detail rather than on the inclusion of behavioral and attitudinal data.

Table 8 summarizes a representative sample of studies that formulated (and in some cases utilized) a life cycle or lifestyle segmentation system. The marketing literature has had a strong interest in life cycle and lifestyle segmentation as early as the 1960s when it was borrowed from the field of sociology [24]. Since then, many researchers have formulated different models. A conceptual and empirical comparison of life cycle models in marketing is presented elsewhere [21]. The transportation community started applying lifestyle and life cycle segmentation concepts in the early 1980s. For example, in 1983 lifestyle and life cycle segmentations were tested against income segmentation for both shopping trip mode choice and destination choice models [20]. The choice models segmented by lifestyles predicted better than the pooled alternative and better than both the life cycle and income segmented models. To the authors' knowledge, a lifestyle segmentation system has not yet

been applied in air travel demand modeling.

4.4 Data

Two data sets are used in this study. The first data set, obtained from a credit-reporting agency, contains a rich set of individual-level demographic and socioeconomic information including lifestyle variables. The second dataset, obtained from an air passenger survey conducted at Hartsfield-Jackson Atlanta International Airport, provides information on the originating and terminating ZIP codes for a sample of resident and nonresident air travelers at the airport. This section describes each data set and known sources of bias as these limitations are important in interpreting the results.

4.4.1 Credit-Reporting Data

U.S. credit-reporting agencies collect information for individuals aged 18 years and older who have a credit history. Unlike census data, credit-reporting data are updated frequently, often on a monthly or quarterly basis. A wide range of data sources is used to populate the credit-reporting database. These sources provide information related to finances (monthly credit card transactions, current mortgage balances, credit scores), socioeconomic and demographic characteristics (number and ages of adults and children in the household, household income, ethnicity, educational level, occupation, home owner or renter, dwelling type, length of time in current residence), and lifestyle preferences and hobbies (interest in travel, fishing, fine arts). The latter lifestyle preferences are obtained from a variety of sources, including product registration forms and specialty credit cards.

The credit-reporting data for this study represent a 10% sample of Atlanta residents (ages 18 years and older) in the 13-county Atlanta metropolitan area, or about 500,000 households. The credit-reporting data reflect characteristics of the Atlanta metropolitan area as of January 2011. Income and lifestyle clusters were the key variables used in the study.

Demographic and socioeconomic variables such as income and age are updated regularly by the credit-reporting agency, and this information may provide a forecasting advantage

Table 8: Summary of Life Cycle and Lifestyle Segmentation in Literature

	Data Source	Location and Year	Number of Households	Number of Clusters	Analysis Method	Application
Marketing						
Wilkes [25]	Consumer Expenditure Survey	U.S. 1988-1990	7,337	15	Categories based on household structure alone	n/a[a]
Brown and Venkatesh [3]	Market research firm (unnamed)	U.S. 1999-2000	746	7	Categories based on household structure and income	Technology adoption model
Du and Kamakura [5]	Panel Study of Income Dynamics	U.S. 1968-2001 (every 4th year)	6,887	13	Categories based on household structure & employment	Allocation of consumption budget Du and Kamakura [6]
Economic Modeling						
Alekander et al. [1]	California PECAS Model Population Synthesizer[b]	California 2000 (base year)	Not specified	Ranges from 12 to 231	Two-step clustering method	PECAS[b,c]
Transportation						
Salomon and Ben-Akiva [20]	Federal Highway Administration	Baltimore, MD 1977	521[d]	5	Cluster analysis (K-means method)	Mode choice and destination choice for shopping trips
Ma and Goulias [14]	Puget Sound Transportation Panel	Seattle, WA 1989-1990	1,376	4	Cluster analysis	Activity participation over time
Krizek and Waddell [12]	Puget Sound Transportation Panel	Seattle, Washington 1997	1,907	9	Factor analysis and cluster analysis	n/a[a]
Krizek [11]	TBI Home Interview Survey	Twin Cities, Minnesota 2001	∼9,000	7	Factor analysis and cluster analysis	n/a[a]
This study	Credit reporting agency	Atlanta, Georgia 2011	426,648	26	Proprietary	Airport passenger allocation model

[a] n/a = not applicable.
[b] PECAS = Production Exchange Consumption Allocation System.
[c] PECAS is a spatial economic model system.
[d] All households in this set were headed by married couples.

over the decennial census, the data source most often used in current air passenger models. Although the exact algorithm the credit-reporting agency uses to populate household income is proprietary, the credit-reporting agency shared some details. Specifically, its income model is recalculated at least quarterly, and its underlying algorithm is rebuilt every few years. Approximately 25% of the income data are obtained from credit applications. Income for the remaining records is imputed using an algorithm that considers data such as age, home ownership, home value, presence of children in the household, occupation, and education.

The credit-reporting agency also uses a confidential algorithm to create lifestyle segmentation variables. Numerous inputs are used to create these segmentation variables, including demographic data, financial data, survey data, transaction data, behavioral and attitudinal data, and trigger data. Trigger data refers to life events (such as the birth of a child) that might cause an individual to move into a different lifestyle cluster. The algorithm segments households into 26 clusters. Each cluster represents a unique lifestyle. A summary table of the 26 clusters based on the U.S. population as of January 2011 is presented in Table 9.

The segmentation model considers consumers' air travel preferences (although the number of annual air trips is not directly observed). The travel information primarily comes from stated preference surveys that ask specific questions related to air travel. For example, one question might read, "Does any member of your household belong to a frequent flyer program?" Individuals who consistently express an interest in air travel across multiple data sources are more likely to appear in the clusters noted in Table 9 that are associated with travel in general or domestic travel, foreign travel, and business travel specifically.

This study is based on 426,648 households. Only observations that contained complete information on household income, age of the head of household, and household lifestyle cluster were included in the analysis. Tables 10 and 11 show the frequency of lifestyle clusters by the income and age categories. Income is used to determine which clusters a household may belong to. For example, a household making $125,000 or more will not appear in clusters 10 to 26. Age is more uniformly distributed across the clusters, although some of the clusters such as "Already Affluent" and "Nice and Easy Grandparents" show

Table 9: Summary of Credit-Reporting Lifestyle Clusters[a]

Cluster			Mean		Travel Indicator			
No.	Name	Percent Composition	HH Income[b] ($1,000)	Head of HH Age[b] (years)	General	Domestic	Foreign	Business
1	Already Affluent	0.3	166	29	x	–	–	–
2	Big Spender Parents	1.2	162	43	x	–	–	–
3	Chic Society	3.7	167	49	x	–	–	–
4	Diamonds-to-Go	5.7	123	48	x	–	–	–
5	Easy Street	0.4	161	64	x	–	–	–
6	Feather-the-Nest	0.4	163	31	x	–	–	–
7	Go-go Families	0.1	166	43	–	–	–	–
8	Home Hoppers	2.4	125	40	–	–	–	–
9	IRA Spenders	5.1	91	67	–	–	–	–
10	Just Sailing Along	7.7	68	31	x	–	x	–
11	Kiddie Kastles	11.9	73	43	–	–	–	–
12	Loose Change	5.3	71	43	–	–	–	–
13	Mid-Life Munchkins	6.1	71	55	x	–	–	x
14	Nice & Easy Grandparents	6.9	68	68	–	–	–	–
15	Oodles of Offspring	2.1	36	28	x	x	–	–
16	Parks, Parts, & Prayers	2.7	31	38	–	–	–	–
17	Quiet Homebodies	9.2	55	43	–	–	–	–
18	Rocky Road	5.3	40	44	–	–	–	–
19	Still Going Strong	1.5	32	63	–	–	–	–
20	Totebaggers	1.4	26	28	x	–	x	–
21	Under-the-Car	1.4	28	37	–	–	–	–
22	Very Spartan	7.2	26	37	–	–	–	–
23	Working Hard	1.8	25	42	–	–	–	–
24	X-tra Needy	2.0	25	66	–	–	–	–
25	Young-at-Heart	3.4	26	70	–	–	–	–
26	Zero Mobility	4.7	25	71	–	–	–	–

[a] Statistics are from the entire U.S. population (not the data used in this research). Table was adapted from one provided by the credit-reporting agency.
[b] HH = household.

distributions skewed towards particular age ranges.

To determine potential biases in the credit-reporting data, age, income, and vehicle ownership characteristics were compared to census data. In general, the credit-reporting data used for this study underrepresent lower-income households, households that do not own a vehicle, and households that rent. These biases are not necessarily a reflection of the credit-reporting database in general but rather reflect the sampling frame the research team used for confidentially linking the credit-reporting data to the Georgia Department of Motor Vehicle's auto ownership database and the Metropolitan Atlanta Rapid Transit Authority's paratransit database. Of the 500,000 records sent to the credit-reporting company, 458,852 records were matched based on the individual's current address. A large portion of the unmatched addresses included those with apartment numbers and those with limited or no credit history. However, in the context of this study, the sampling bias may have minimal impact on the results because people who frequently make air trips tend to be mid- to high-income individuals who have mature credit reports.

4.4.2 Airport Passenger Survey Data

The secondary data used in this research are from a departing passenger survey conducted by Hartsfield Planning Collaborative at Hartsfield-Jackson Atlanta International Airport for a peak week in July 2009. The survey collected information for 12,075 departing individuals, of whom 5,037 were originating passengers and 7,038 were connecting passengers. The survey was a random stratified single-stage cluster sample. The passengers chosen to participate were identified by randomly selecting flights over a two-week survey period from four mutually exclusive sample groups of flights (hub airline domestic departures, other airline domestic departures, hub airline international departure, and other airline international departure). Each passenger on the selected flights was considered an elementary unit of a cluster (i.e., the flight) and was asked to complete a questionnaire. About 63% of the passengers on the surveyed flights completed a questionnaire [8].

The airport planning department provided 2,456 records of the originating passengers, or about 49% of the surveyed originating passengers. Among these records, 1,759 represented

Table 10: Frequency of Lifestyle Clusters Versus Household Income

Cluster	Household Income												
	$0-$14,999	$15,000-$19,999	$20,000-$29,999	$30,000-$39,999	$40,000-$49,999	$50,000-$74,999	$75,000-$99,999	$100,000-$124,999	$125,000-$149,999	$150,000-$174,999	$175,000-$199,999	$200,000-$249,999	$250,000+
Already Affluent	0	0	0	0	0	0	0	0	454	238	96	35	21
Big Spender Parents	0	0	0	0	0	0	0	0	2,669	824	304	210	175
Chic Society	0	0	0	0	0	0	0	0	7,339	2,713	1,205	871	624
Diamonds-to-Go	0	0	0	0	3,030	7,526	4,865	2,982	13,963	5,206	2,389	1,993	1,798
Easy Street	0	0	0	0	0	0	0	0	912	267	113	87	73
Feather-the-Nest	0	0	0	0	0	0	0	0	555	217	97	53	39
Go-go Families	0	0	0	0	0	0	0	0	89	26	11	4	9
Home Hoppers	0	0	0	0	1,488	2,633	1,476	810	421	71	16	6	16
IRA Spenders	0	0	0	0	4,384	10,549	6,788	3,943	969	294	153	83	73
Just Sailing Along	0	0	0	0	6,313	8,767	3,932	1,496	0	0	0	0	0
Kiddie Kastles	0	0	0	0	8,905	25,565	20,019	12,645	0	0	0	0	0
Loose Change	0	0	0	0	4,354	7,442	3,737	1,836	0	0	0	0	0
Mid-Life Munchkins	0	0	0	0	6,566	16,965	11,359	6,739	0	0	0	0	0
Nice & Easy Grandparents	0	0	0	0	4,587	7,372	3,196	1,527	0	0	0	0	0
Oodles of Offspring	861	610	1,758	2,434	420	403	134	43	0	0	0	0	0
Parks, Parts, & Prayers	997	1,014	3,465	5,201	280	266	90	35	0	0	0	0	0
Quiet Homebodies	1,163	873	2,980	5,889	9,462	14,282	5,418	2,192	0	0	0	0	0
Rocky Road	1,527	1,303	3,949	4,723	1,213	1,435	710	208	0	0	0	0	0
Still Going Strong	1,132	761	1,963	2,825	133	247	79	44	0	0	0	0	0
Totebaggers	1,699	1,209	2,546	1,877	0	0	0	0	0	0	0	0	0
Under-the-Car	581	677	2,300	3,757	0	0	0	0	0	0	0	0	0
Very Spartan	5,730	4,402	11,095	10,280	0	0	0	0	0	0	0	0	0
Working Hard	1,363	790	1,703	1,664	49	43	18	12	0	0	0	0	0
X-tra Needy	1,716	1,208	2,840	2,891	0	0	0	0	0	0	0	0	0
Young-at-Heart	2,045	1,607	4,181	5,628	0	0	0	0	0	0	0	0	0
Zero Mobility	1,609	1,092	2,715	3,222	0	0	0	0	0	0	0	0	0

Table 11: Frequency of Lifestyle Clusters Versus Head of Household Age

	Age (years)							
Cluster	18-29	30-39	40-49	50-59	60-69	70-79	80-89	90+
Already Affluent	206	568	1	3	1	0	0	0
Big Spender Parents	4	628	2,250	1,044	12	2	1	0
Chic Society	67	1,308	3,296	3,204	2,794	1,007	231	46
Diamonds-to-Go	13	5,396	17,279	14,016	4,430	857	98	27
Easy Street	2	11	19	633	580	115	15	4
Feather-the-Nest	184	671	1	28	23	3	1	0
Go-go Families	1	41	61	30	0	0	0	0
Home Hoppers	983	3,020	905	691	434	169	75	31
IRA Spenders	2	38	105	5,863	11,465	6,448	1,859	336
Just Sailing Along	4,608	6,107	989	408	19	9	6	5
Kiddie Kastles	899	13,994	32,069	15,905	184	89	22	9
Loose Change	518	3,621	6,628	3,155	68	43	16	5
Mid-Life Munchkins	2,013	3,612	220	12,389	15,038	5,367	662	107
Nice & Easy Grandparents	4	52	108	4,506	5,898	2,661	892	279
Oodles of Offspring	2,276	2,520	24	20	8	3	1	0
Parks, Parts, & Prayers	1,358	2,533	2,412	953	13	6	6	0
Quiet Homebodies	376	6,607	13,934	7,381	1,102	351	57	24
Rocky Road	19	2,398	4,661	2,446	298	121	65	13
Still Going Strong	2	199	416	1,889	2,664	1,386	203	40
Totebaggers	1,061	1,015	26	18	11	7	5	1
Under-the-Car	788	1,772	1,893	1,290	37	22	14	2
Very Spartan	1,609	3,911	4,913	2,093	48	32	11	0
Working Hard	477	771	659	847	688	245	48	12
X-tra Needy	11	330	700	2,005	2,003	944	339	115
Young-at-Heart	2	21	36	2,585	4,575	3,752	1,389	263
Zero Mobility	2	24	33	1,918	2,938	1,762	800	219

trips from private residences and 1,599 of those had corresponding ZIP code information. Trips from places of business or hotels were excluded because the credit-reporting data provide household information, not workplace information. This study focuses on predicting home-based trips in the 13-county metropolitan area defined as those trips that have their non-airport trip end at a private residence. The number of surveys with their non-airport trip end at a private residence represents 72.3% of trips in the peak week survey [8], and therefore the model covers a significant portion of airport passenger trips without considering workplace-based trips to the airport.

Because the catchment area of Hartsfield-Jackson is larger than the 13-county metropolitan area in the credit-reporting data, the ZIP codes that fall within the 13-county needed to be identified. ZIP code and county boundaries do not necessary coincide. QGIS, an open source geographic information system [16], was used to calculate each ZIP codes centroid. A total of 143 ZIP codes whose centroids fell within any of the 13 counties were retained in the sample. The total sample from the air passenger survey data used in this analysis, then, was 1,131 cases within the 13-county metropolitan area.

4.5 Methodology

The objective of this study was to compare a least-squares regression model based on the functional form used in the existing ARC airport passenger model with one that uses lifestyle segmentation variables. However, limitations associated with the credit-reporting and air travel survey databases posed two key methodological challenges. The first challenge was in defining an appropriate trip rate from the air passenger survey that could be linked to the credit-reporting data at the ZIP code level. The second challenge was identifying and eliminating outliers. This section describes the least-squares regression models in general and how these two methodological challenges were addressed.

As noted in the literature on traditional demographic models in air travel [22], a simple linear model is one of the models used most often. This takes the following form:

$$Y = \alpha + \beta X + \epsilon \tag{4}$$

where Y is the dependent variable, X is a matrix of independent variables, α is a constant, β is a matrix of the coefficients describing how a change in each X affects Y, and ϵ is a random error term with mean zero.

4.5.1 Independent Variables

The credit-reporting and air passenger survey data were linked at the ZIP code level. The credit-reporting data were used to create the independent variables, and the air passenger survey was used to create the dependent variable. The first set of regression models replicated the independent variables used in the existing ARC airport passenger model, which was based solely on income group populations in each TAZ from the census data. To make the replicated models most similar, the same income groups were formulated with the credit-reporting data and summarized at the ZIP code level. In contrast to the ARC model, the income group populations were standardized by the total population in each ZIP code. For example, the value of x for high-income households in ZIP code 30308 is expressed as a percentage:

$$x_{H(30308)} = \frac{N_{H(30308)}}{T_{(30308)}} * 100 \tag{5}$$

where $N_{H(30308)}$ is the number of households with an income greater than \$100,000 in ZIP code 30308 and $T_{(30308)}$ is the total number of households in ZIP code 30308 in the crediting-reporting data.

The second set of regression models, the lifestyle models, used the lifestyle variables. For example, the percentage for the "Already Affluent" cluster in ZIP code 30308 is:

$$x_{1(30308)} = \frac{N_{1(30308)}}{T_{(30308)}} * 100 \tag{6}$$

where $N_{1(30308)}$ is the number of households in the "Already Affluent" cluster in ZIP code 30308 and $T_{(30308)}$ is the total number of households in ZIP code 30308 in the credit-reporting data.

4.5.2 Dependent Variables

The dependent variable is a measure of the average number of air trips at the ZIP code level. It was necessary to create an average pseudo trip rate for each ZIP code because the air passenger survey did not collect information about individuals' annual air travel rates. The average pseudo trip rate, y, was calculated by summing the number of surveyed individuals according to the home ZIP code and dividing by the total number of households (as determined in the credit-reporting database):

$$y_{(30308)} = \frac{S_{(30308)}}{T_{(30308)}} * 1000 \tag{7}$$

where $S_{(30308)}$ is the number of survey respondents in the air passenger data with home ZIP code 30308 and $T_{(30308)}$ is the total number of households in ZIP code 30308 in the credit-reporting data. Although the survey asked individuals if they were traveling together, that information was not provided to the researchers. Thus, the pseudo trip measure used in this analysis does not control for multiple responses from a given party. The number of survey respondents per ZIP code ranged between 0 and 33 with a median of 6.5 and mean of about 8. Because the number of survey respondents in each ZIP code was small compared with the sample in the credit-reporting data, the trip ratio was multiplied by 1,000, which makes the model results easier to interpret. This assumption, although likely biased in an immeasurable way, should be viewed as a proxy measure for the number of air trips per household for each ZIP code. The ARC model's dependent variable had two differences: it was not standardized by population and the survey numbers were projected up to match the total number of enplanements as reported by the FAA. The authors chose not to do the latter step because it did not provide any added benefit for comparing the models. The dependent variable is referred to simply as average trips per household for the remainder of the paper. For each ZIP code:

$$y_{(ZIP)} = \alpha + \beta_i x_{i(ZIP)} \tag{8}$$

where i indexes the independent variables included in a given model (e.g., x_H or x_1).

4.5.3 Outliers

The free software environment for statistical computing and graphics called R was used to analyze the data and models [17]. Several successive outlier tests using the package car [7] indicated a large number of outliers in the data. Both sets of regression models (the ones based on income groups and those based on lifestyle clusters) presented the same severe outliers. These outliers resulted from the average trips per household by ZIP code, which had a distribution that was largely skewed to the right. These outliers could be due to the errors associated with the large difference in sample sizes between the credit-reporting data and the air passenger survey. Furthermore, there were obvious errors in a few of the ZIP codes where the number of surveys conducted in the air passenger data was high with no existing residential population and therefore no cases in the credit-reporting data. For example, the airport ZIP code was reported as the home ZIP code seven times in the air passenger survey but was not present in the credit-reporting data because no residences are associated with that ZIP code. To remove these and similar survey coding errors, nine ZIP codes that had an average trip frequency measure greater than ten trips per household were removed, leaving 134 ZIP codes in the data.

4.5.4 Model Selection

To compare different model specifications, the adjusted R-squared statistic was used as the primary reported measure of fit. However, additional diagnostics were examined to ensure that a linear regression model was appropriate. Some of these included plots of residuals versus fitted values (which were regularly distributed about zero), the normal Q-Q plots (which produced an approximate straight line), and the standardized residuals versus leverage plots (which showed that all of the points fell within a reasonable Cook's distance of one). A normal Q-Q plot is a graphical method for comparing the normal probability distribution with a given set of data—in this case, the standardized residuals—by plotting their quantiles against each other. The linearity of points on this type of plot suggests that the residuals are normally distributed. Cook's distance is a measure of the effect of a given observation on the regression result. There are different opinions about what cutoff values

of Cook's distance indicate an outlier. For a standardized residuals versus leverage plot, distances larger than one suggest the presence of a possible outlier or a poor model. Refer to Neter et al. [15] for further descriptions of Q-Q plots and Cook's distance.

Two traditional models based on income groups and numerous models based on the lifestyle clusters were estimated. The leaps package [13] in R was used to select the optimal group of lifestyle clusters to retain in models that were built parsimoniously. Specifically, the regsubsets command associated with the leaps package was used to identify the subset of clusters that provided the best adjusted R-squared value. To clarify, the command allows the user to specify the maximum number of desired variables in the model, and then the regsubsets routine determines which of the available variables maximizes the adjusted R-squared value for each valid model size. By iteratively estimating the best subsets returned by the regsubsets routine for each model size, two preferred models were selected: one containing three clusters and another containing five clusters.

4.6 Results

Table 19 presents the results for four regression models. Models 1 and 2 are the traditional models that use income to emulate the ARC's airport passenger model, and Models 3 and 4 use lifestyle clusters. Because of the data limitations discussed previously (particularly the small number of observations available from the air passenger survey), the results should be interpreted as directional evidence, not as absolute trip frequency predictions. In general, the lifestyle clusters predict the average number of air passenger trips better than income groups.

Model 1, which only includes the percent of high-income households in the ZIP code, has the lowest adjusted R-squared value. Model 2 includes three of the income groups. Although both the medium-high and medium-low income in Model 2 are not statistically significant at the 95% confidence level, the model fit does improve slightly. The counter-initiative results of the β estimates in Model 2 reflect the same nuances captured in the ARC's air passenger model; ARC constrained the β estimates shown in Figure 8 to ensure that they were increasing with income [23]. For the purposes of this comparison, Model 2

Table 12: Regression Models Predicting Psuedo Trip Rate

	Model 1		Model 2		Model 3		Model 4	
	Estimate	t-Statistic	Estimate	t-Statistic	Estimate	t-Statistic	Estimate	t-Statistic
Intercept (α)	1.5259	7.586	2.9651	3.485	1.6316	5.291	0.9094	2.524
Travel cluster (β)								
Chic Society, X_3	–	–	–	–	–	–	0.0881	2.161
Diamonds-to-Go, X_4	–	–	–	–	0.1457	6.656	0.1311	4.832
Just Sailing Along, X_{10}	–	–	–	–	0.0756	3.459	0.0866	3.547
Totebaggers, X_{20}	–	–	–	–	–	–	0.1843	3.392
Nontravel cluster (β)								
Kiddie Kastles, X_{11}	–	–	–	–	-0.0564	-3.644	–	–
Nice & Easy Grandparents, X_{14}	–	–	–	–	–	–	-0.1650	-2.218
Income group (β)								
High	0.0580	7.213	0.0468	3.739	–	–	–	–
Medium-high	–	–	*-0.0163*	*-1.688*	–	–	–	–
Medium-low	–	–	*-0.0183*	*-1.371*	–	–	–	–
Low	–	–	–	–	–	–	–	–
Model statistics								
Adjusted R-squared	0.2773		0.2854		0.4206		0.4635	

Italics denote insignificant parameters at 95% confidence level; dashes denote a variable that was not estimated in the particular model.

Table 13: Qualitative Description of Selected Clusters

Cluster	Demographic Profile	Interests
Chic Society	Few kids, high home ownership and values, high education, mail responsive	Stocks/bonds, apparel, charities, fitness, cultural events, antiques, fashion, travel, multiple credit cards
Diamonds-to-Go	Home owners, high home values, kids, white collar, mail responsive	Home furnishings, stocks, computers, gourmet cooking, gardening, travel, multiple credit cards
Just Sailing Along	No children, renters, white collar, students, short length of residence	Camping equipment, electronics, wines, gourmet food, new technology, outdoor activities, travel abroad
Totebaggers	Young, single adults, no children, high mobility, students, sales/service	New technology, personal computers, electronics, fitness, many sports, cultural events, travel abroad
Kiddie Kastles	Homeowners, children in household, white collar, college graduates, mail responsive	Computers, video cameras, kids items, fitness, outdoor activities, automotive work, multiple credit cards
Nice & Easy Grandparents	Empty nesters, homeowners, long lengths of residence, retired or white collar, high education	Tools, audio equipment, stocks, fundraising, gardening, golf, crafts, civic and bible activities, grandchildren

Table is adapted from one provided by the credit-reporting agency.

was left unconstrained to demonstrate its true fit. Model 1 was included as an alternative traditional model that intuitively makes sense without constraints.

Models 3 and 4 both have one cluster from the set of non-travel clusters and multiple clusters from the set of travel clusters as defined in Table 9. These clusters are described qualitatively in Table 13. In each of the models, the variables associated with the non-travel clusters (i.e., "Kiddie Kastles" and "Nice and Easy Grandparents") have negative β estimates, and the variables associated with the travel clusters (i.e., "Chic Society," "Diamonds-to-Go," "Just Sailing Along," and "Totebaggers") have positive β estimates. Each of the lifestyle clusters is significant at the 99% confidence level. Models 3 and 4 have significantly higher adjusted R-squared values than Models 1 and 2.

To interpret the limits of Models 3 and 4, Figure 9 and Tables 14 and 15 are shown. Figure 9 shows a boxplot for each of the lifestyle cluster variables used in Models 3 and 4. A boxplot is a graphic representation of the five-number summary of numerical data: minimum value, lower quartile, median, upper quartile, and maximum value. The shaded region shows the interquartile range from the first quartile to the third (i.e., from the 25th

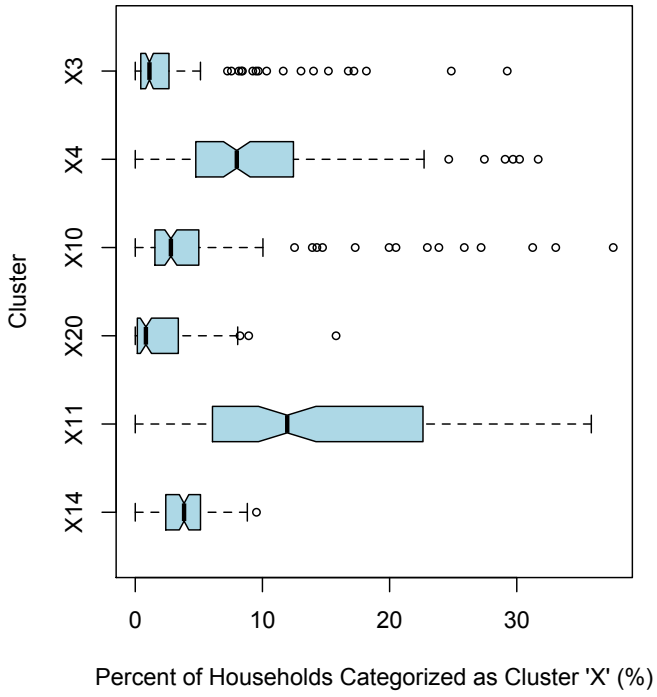


Figure 9: Boxplots showing range of values for each lifestyle cluster in Models 3 and 4.

to the 75th percentile). The middle line marks the median. The whiskers, shown as a dotted line, go from the minimum to the maximum value. If the distance from the minimum value to the first quartile or from the maximum value to the third quartile is more than one and a half times the interquartile range, then the values are considered outliers and are denoted by a circle. The boxplots for each of the lifestyle cluster variables show that the scope of Models 3 and 4 include percentages up to about 35% for the lifestyle cluster composition in each ZIP code. In other words, the model can only predict with accuracy an average trip rate for ZIP codes that have 35% or less of their population in a particular lifestyle cluster.

Tables 14 and 15 present observed and hypothetical cases ordered by their fitted Y-values, $\hat{Y}$, for Models 3 and 4, respectively. The observed cases, which are denoted by a real ZIP code listed in Tables 14 and 15, show the actual percent compositions for each of the lifestyle clusters and $\hat{Y}$. These cases were selected from the 134 observed ZIP codes because their $\hat{Y}$ values are closest to the minimum, median, mean, and maximum values of

Table 14: Real and Hypothetical Cases for Model 3 Ordered by Fitted Values of Pseudo Trip Rate

ZIP Code	Case Type	X_4	X_{10}	X_{11}		$\hat{Y}$
	$1.63 + 0.14X_4 + 0.08X_{10} - 0.06X_{11} = \hat{Y}$					
30238	Real	2.4	0.5	24.0	0.7	(Min)
–	*Min*	*0*	*0*	*0*	*1.6*	–
30019	Real	13.2	2.7	31.1	2.0	(Median)
–	*Median*	*8.0*	*2.8*	*12.0*	*2.3*	–
–	*Mean*	*9.2*	*4.9*	*14.3*	*2.5*	–
30066	Real	13.3	2.7	22.1	2.5	(Mean)
30327	Real	31.7	3.5	1.8	6.4	(Max)
–	*Max*[a]	*31.7*	*37.6*	*35.9*	*7.1*	–

Italics denote hypothetical cases, which are based on the summary statistics of each lifestyle cluster; min = minimum; max = maximum.

[a] This hypothetical case, where each of the cluster percentages is its maximum value, is technically outside the scope of the model and is also impossible because the percentages add up to more than 100. It is still helpful though because it shows that even in this case, the predicted pseudo trip rate is reasonable compared with the minimum observed case (ZIP code 30238).

Table 15: Real and Hypothetical Cases for Model 4 Ordered by Fitted Values of Pseudo Trip Rate

ZIP Code	Case Type	X_3	X_4	X_{10}	X_{20}	X_{14}		$\hat{Y}$
	$0.91 + 0.09X_3 + 0.13X_4 + 0.09X_{10} + 0.18X_{20} - 0.17X_{14} = \hat{Y}$							
30069	Real/*Min*	0	0	0	0	0	0.9	(Min)
–	*Median*	*1.1*	*8.0*	*2.8*	*0.8*	*3.8*	*1.8*	–
30168	Real	0.1	4.2	1.3	3.8	1.3	2.1	(Median)
–	*Mean*	*2.6*	*9.2*	*4.9*	*2.0*	*3.8*	*2.5*	–
30033	Real	4.0	12.2	6.5	2.7	8.8	2.5	(Mean)
30327	Real	29.2	31.7	3.5	0.2	2.2	7.6	(Max)
–	*Max*[a]	*29.3*	*31.7*	*37.6*	*15.8*	*9.5*	*12.2*	–

Italics denote hypothetical cases, which are based on the summary statistics of each lifestyle cluster; min = minimum; max = maximum.

[a] This hypothetical case, where each of the cluster percentages is its maximum value, is technically outside the scope of the model and is impossible because the percentages add up to more than 100. It is still helpful though because it shows that even in this case, the predicted pseudo trip rate is reasonable compared with the minimum observed case (ZIP code 30069).

$\hat{Y}$. The hypothetical cases, shown in italics, are derived from the summary statistics for each lifestyle cluster used in the model—that is, the minimum, median, mean, and maximum of each of the independent variables.

For Model 3, the maximum $\hat{Y}$ is about 6.5 trips per household and the minimum is about 1 (Table 14). For Model 4, the maximum $\hat{Y}$ is about 7.5 and the minimum is also about 1 (Table 15). Both models have medians and means falling between 2 and 2.5 trips, which is reasonable. Given the hypothetical case where all the variables are their maximum percentage values (which is not technically possible because they add up to more than 100%), the $\hat{Y}$ values for each of these hypothetical cases remain at a reasonable 7 and 12 trips, respectively. These values demonstrate that the two regression models are predicting reasonably in addition to having significantly improved adjusted R-squared values.

4.7 Future Research

There are many opportunities to improve aviation forecasting models by incorporating richer demographic and socioeconomic information, particularly lifestyle segmentation information. It would be interesting to repeat this study with an airport passenger survey that included the total number of annual air trips taken by the survey respondents as well as the residential and work street addresses. This approach would help overcome a key methodological challenges encountered in this study, namely, the need to create an average pseudo trip rate at a high level of geographic aggregation. A more accurate measure of trip rates calculated for smaller geographic areas, such as TAZs or block groups, could also help uncover idiosyncrasies that exist across TAZs but when aggregated do not appear at the ZIP level. The work street addresses would be helpful in separately predicting those trips to the airport that are made directly from work.

It would also be interesting to repeat this study for different metropolitan areas to see if the estimated models are transferrable to other regions. The credit-reporting database is unique in that it maintains consistent definitions of clusters across the United States. This factor is important because one of the main arguments against using lifestyle variables is that they are dependent on the data used and therefore are not transferable to different

regions when the data source changes. This type of analysis would contribute to the lifestyle literature in general because it would look at the transferability of the lifestyle clusters across the U.S. in addition to the transferability of the models specified in this study.

To use the lifestyle cluster models for forecasting future trips to the airport specifically, a model would need to be developed to predict changes in lifestyle percent compositions for a desired geographic aggregation level (e.g., ZIP code, TAZ). Forecasting could be accomplished by tracking, through the credit-reporting database, how individuals change lifestyle clusters over time.

4.8 Conclusions and Policy Implications

This study provides evidence that non-traditional data sources can be used in air travel forecasting applications. Regression models based on lifestyle clusters from a non-traditional data source exhibited much higher adjusted R-squared values than did regression models based on income. These results suggest that alternative data sources—namely, those that provide information about consumer preferences and attitudes revealed through lifestyle segmentation—can improve the forecasting accuracy of airport passenger models.

Non-traditional data sources may have other advantages. For example, the credit-reporting data used in this study can be obtained at a lower cost than some traditional sources, such as two-day travel surveys. Further, the credit-reporting data are updated more frequently than air passenger and census data. Monthly or quarterly updates of household income, employment status, credit balances, and lifecycle clusters could be valuable in analyzing how a particular economic or political shock could impact the aviation industry [4]. With the recent instability of the economy, it has become increasingly important to examine the impact of these types of shocks.

The data used in this study, particularly the lifestyle variables, are often purchased by firms that want to target a customer segment (e.g., by offering promotions for particular products or services). It would be interesting to see if a similar approach could be used to market existing or new airport services. For example, it may be possible to increase the number of trips to the airport taken using Metropolitan Atlanta Rapid Transit Authority

by identifying environmentally conscious air travelers through the lifestyle variables and sending them information on park-and-ride lots and train schedules for stations near their residences.

4.9 Acknowledgements

This research was supported by the 2010-2011 Graduate Research Award Program on Public-Sector Aviation Issues. Part of this research was also supported by a National Science Foundation (NS) Graduate Research Fellowship and NSF CAREER grant. The authors thank the program mentors, including Dipasis Bhadra, Michael Drollinger, Geoffrey Gosling, and Annalisa Weigel, for the beneficial comments and feedback, along with Shelley Lamar from HartsfieldJackson International Airport for her efforts in data compilation. The authors also thank program manager, Lawrence Goldstein, for his help.

4.10 References

[1] B. Alekander, D. D. Silva, J. E. Abraham, J. D. Hunt, and S. Gao. "Lifestyle Clusters for Labor Force Participation, Occupation, and Housing Use in Spatial Economic Modeling." In *12th International Conference on Computers in Urban Planning and Urban Management*, 2011.

[2] M. E. Ben-Akiva and S. R. Lerman. "A Behavioral Analysis of Automobile Ownership and Modes of Travel Publication DOT-OS-3005603." Technical report, Cambridge Systematics, Inc., 1976.

[3] S. A. Brown and V. Venkatesh. "Model of Adoption of Technology in Households: A Baseline Model Test and Extension Incorporating Household Life Cycle." *MIS Quarterly*, 29(3):399–426, 2005.

[4] J. P. Cripwell and G. Gosling. "ACRP Problem No. 10-03-08: Influence of Demographics and Socio-Economic Factors on Air Travel Demand." Technical report, Transportation Research Board of the National Academies, Washington, D.C., 2009.

[5] R. Y. Du and W. A. Kamakura. "Household Life Cycles and Lifestyles in the United States." *Journal of Marketing Research*, 43(1):121–132, 2006.

[6] R. Y. Du and W. A. Kamakura. "Where Did All That Money Go? Understanding How Consumers Allocate Their Consumption Budget." *Journal of Marketing*, 72(6): 109–131, 2008.

[7] J. Fox and S. Weisberg. *An R Companion to Applied Regression.* Thousand Oaks CA: Sage, 2nd edition, 2011.

[8] Hartsfield Planning Collaborative. "Peak Week Survey Results." Technical report, 2009.

[9] R. Kitamura. "Life-Style and Travel Demand." *Transportation*, 36:679–710, 2009.

[10] J. D. Kressner and L. A. Garrow. "Lifestyle Segmentation Variables as Predictors of Home-Based Trips for Atlanta, Georgia Airport." *Transportation Research Record: Journal of the Transportation Research Board*, 2266:20–30, 2012.

[11] K. J. Krizek. "Lifestyles, Residential Location Decisions, and Pedestrian and Transit Activity." *Transportation Research Record: Journal of the Transportation Research Board*, 1981:171–178, 2006.

[12] K. J. Krizek and P. Waddell. "Analysis of Lifestyle Choices: Neighborhood Type, Travel Patterns, and Activity Participation." *Transportation Research Record: Journal of the Transportation Research Board*, 1807:119–128, 2002.

[13] T. Lumley. "Leaps: Regression Subset Selection. R package. Version 2.9." `http://CRAN.R-project.org/package=leaps`, 2009. Uses Fortran code by Alan Miller.

[14] J. Ma and K. G. Goulias. "A Dynamic Analysis of Person and Household Activity and Travel Patterns Using Data from the First Two Waves in the Puget Sound Transportation Panel." *Transportation*, 24(3):309–331, 1997.

[15] J. Neter, M. Kutner, C. J. Nachtscheim, and W. Wasserman. *Applied Linear Statistical Models, 5th Edition*. McGraw-Hill, 2004.

[16] OSGeo Project, GNU Public License. "Quantum GIS: About QGIS." `http://www.qgis.org/en/about-qgis.html`, 2011. Accessed on 13 July 2011.

[17] R Development Core Team. "R: A Language and Environment for Statistical Computing." `http://www.R-project.org`, 2011.

[18] S. Reichman. "Travel Adjustments and Life Styles: A Behavioral Approach." In P. R. Stopher and A. H. Meyburg, editors, *Behavioral Travel-Demand Models*, pages 143–152, Lexington, Mass., 1975. Lexington Books.

[19] S. Reichman. "Instrumental and Life-Style Aspects of Urban Travel Behavior." *Transportation Research Record: Journal of the Transportation Research Board*, 649:3842, 1977.

[20] I. Salomon and M. E. Ben-Akiva. "The Use of the Life-Style Concept in Travel Demand Models." *Environment and Planning A*, 15:623–638, 1983.

[21] C. M. Schaninger and W. D. Danko. "A Conceptual and Empirical Comparison of Alternative Household Life Cycle Models." *Journal of Consumer Research*, 19:580–594, 1993.

[22] W. Spitz and R. Golaszewski. "ACRP Synthesis 2: Airport Aviation Activity Forecasting: A Synthesis of Airport Practice." Technical report, Transportation Research Board of the National Academies, Washington, D.C., 2007.

[23] Unknown. "The Travel Forecasting Model Set For the Atlanta Region 2010 Documentation." Technical report, Atlanta Regional Commission, 2000.

[24] W. D. Wells and G. Gubar. "Life Cycle Concept in Marketing Research." *Journal of Marketing Research*, 3(4):355–363, 1966.

[25] R. E. Wilkes. "Household Life-Cycle Stages, Transitions, and Product Expenditures." *Journal of Consumer Research*, 22(1):27, 1995.

CHAPTER V

RESIDENTIAL LOCATION CHOICE MODEL

J. D. Kressner and L. A. Garrow. "Leveraging targeted marketing data in travel demand modeling: An application in residential location choice modeling.". Georgia Institute of Technology. Working paper, 2014

5.1 Abstract

The purpose of this research is to test that non-traditional data sources, specifically third-party targeted marketing data and mobile phone data, can be used as a supplement to or potential replacement for traditional household travel surveys. The scope of this paper is limited to testing with a simplified multinomial logit (MNL) residential location choice model. Three models were estimated: (1) a base-line model using variables commonly found in residential location choice models; (2) a model that adds a unique measure, derived from the targeted marketing data, of a household's tendency to choose locations with households similar to their own; and (3) a model that adds an analogous measure of a household's tendency to choose locations that are without households most opposite to their own. The sensitivity of the third model to the sample size and the number of random alternatives included in the model estimation is tested using Monte Carlo simulations. The model results indicate that residential location choice models can be estimated using non-traditional data sources, and in fact, model fit is better due to the availability of more variables than those currently included in household travel surveys. The Monte Carlo experiment indicates that parameter estimates are quite sensitive to household sample size. With sample sizes commonly used in the literature (N=500-1,000), it was shown that the standard deviation of parameter estimates was relatively high when compared to the "true" parameter estimates, often producing parameter estimates with incorrect signs. The large sample sizes and additional variables available with targeted marketing and mobile phone data offer several cost and modeling benefits over traditional surveys for residential location modeling.

5.2 Introduction

Residential location choice modeling has a rich history. Urban and transportation planners recognized early on that residential location decisions determine how households connect to the urban environment, thereby helping to shape transportation infrastructure, land-use policies, and urban form. In practice, residential location choice models are often a part of an integrated model of land-use and transportation. Some examples include PECAS [16], UrbanSim [45], DELTA [8], and METROPILUS [36]. Residential location choice modeling has always presented methodological challenges to researchers. Some of these challenges include handling the large number of residential choices available through different sampling strategies [14, 26, 27, 38], addressing price endogeneity [13] and other endogenous long-term and short-term choices like work location and vehicle ownership that influence the residential location decision [9, 18, 35, 37, 47], accounting for self-selection effects (see [30] for a review of recent approaches), and incorporating spatial correlation across choice alternatives and decision-makers [4, 39]. Researchers continue to advance and refine methods for addressing these methodological challenges.

A fact that is often overlooked, though, is that to successfully estimate and implement these advanced models, larger and more detailed estimation samples are needed. However, in an era of decreasing budgets, infrequent collection of household travel surveys, and declining survey response rates, it is often difficult for researchers to obtain sample sizes that would allow them to specify robust models. Additionally, nonresponse and sampling biases of household travels surveys are generally worsening. For example, a large portion of our household travel surveys, including the most recent National Household Travel Survey [10], draw samples entirely from the set of households with fixed landline telephones. Yet, the International Telecommunication Union (ITU) [19] reports that only 44.0 fixed phone subscriptions existed per 100 inhabitants nationally in 2012, which includes subscriptions for businesses. For comparison, 98.2 mobile phone subscriptions existed per 100 inhabitants in 2012, which according to the Pew Research Center translates to mobile phone ownership by 91% of all adults in the U.S. [20, 34].

Issues related to the declining quality of household travel surveys and their impact on infrastructure and policy decisions have been extensively discussed within the transportation community [40, 41, 44, 48]. Multiple advancements have been proposed including GPS-based surveys to improve trip reporting and location accuracy, advanced computer-assisted telephone interviewing (CATI) programming to ease respondent burden and increase response rates, and more detailed residence and workplace tabulations at the traffic analysis zone (TAZ) level from the Census Transportation Planning Package (CTPP) to improve the expansion of survey results to a region. However, none of these suggestions address the root of the problem: interview-based surveys. The proposed solutions continue to suffer from the high cost and difficult recruitment process of surveys. In this paper, we consider an approach to obtaining passively-collected data. We explore the use of targeted marketing data, sometimes also referred to simply as marketing or consumer data. Targeted marketing data are collected by independent companies who make their detailed demographic and socioeconomic information at the household- and person-level available for purchase.

In this work, we estimate a simple residential location choice model using targeted marketing, mobile phone, and U.S. Census data. The model is subsequently expanded to include additional variables particular to the targeted marketing data. Lastly, we test the sensitivity of the model to the size of the number of sampled households and the number of sampled alternatives included in the model estimation using Monte Carlo simulations.

5.3 Data

Residential location choice models typically model location decisions as a function of neighborhood attributes (such as number of housing units), accessibility variables (such as travel time to work), and interactions among household and alternative characteristics. We use three types of data: (1) targeted marketing data for household characteristics and neighborhood attributes, (2) mobile phone data for neighborhood attributes, and (3) U.S. Census data for accessibility variables and neighborhood attributes. Table 16 summarizes the data source for each variable included in the specification. We use a specification that is similar to one used in another study [47], but due to differences in available information some

Table 16: Data Source for Each Variable in the Model Specification

Variable	Source
Neighborhood Attributes	
log(Average Income)	Targeted marketing data
log(Population Density)	2010 Decennial Census
log(Number of Housing Units)	2010 Decennial Census
Accessibility Variables	
Avg Commute Time x *log*(Employment Density)	ACS[a] x Mobile phone data
Interactions of Household and Alternative Characteristics	
Household Size x Avg Household Size	Targeted marketing data x 2010 Decennial Census
Household Income x Avg CMV[b]	Targeted marketing data
Lifestyle Similarity Measure	Targeted marketing data
Lifestyle Dissimilarity Measure	Targeted marketing data

[a] ACS = American Community Survey 5-year estimates 2006-2010
[b] CMV = current market value

substitutions are made.

5.3.1 Targeted Marketing Data

Targeted marketing firms compile information about individuals and households from a variety of sources such as public records, credit reports, credit card transactions, email lists, and internet behavior. They aim to include all individuals aged 18 years or older. The targeted marketing firms sell these data to companies wanting to customize marketing campaigns to potential customers. Coincidentally, these data contain the majority of household and individual demographic and socioeconomic fields that are used in travel demand forecasting applications. Targeted marketing data have been used in several prior studies relevant to travel demand modeling [7, 22, 23, 25].

We use household-specific information including household income and household size from the targeted marketing data in this study. We also use the data to calculate alternative characteristics including the average income and the average current marketing value of housing. Additionally, we use a lifestyle clustering system from the targeted marketing data, which is described in more detail in the next section.

5.3.1.1 Lifestyle Clustering

The targeted marketing data contain information about individuals' behavioral preferences and attitudes that is captured in a "lifestyle segmentation" variable. Lifestyle clusters are commonly used in targeted marketing applications as a way to identify individuals who are more likely to be interested in particular products. These lifestyle clusters incorporate information about household income, ages of individuals in the household, the types of products the household typically purchases, and changes in household structure (such as the birth of a child).

For this study, we use a segmenting system developed by the targeted marketing firm that classifies households into 26 clusters, each representing a unique lifestyle. The cleverly-named clusters range from the young and wealthy "Already Affluent" to the least prosperous "Zero Mobility." As an example, the "Easy Street" cluster is described like this: "The households in this niche are typically older, white collar and educated. They have grown children, possibly still living with them. All of the households within this niche own their homes and have lived at the same address for 7 years or more. On average, their homes are worth about $250,000. They are more likely than the general population to have a pool and to own a vacation home." These lifestyle clusters, as well as the individual behavioral variables used to classify them, provide an opportunity to incorporate behavioral preferences and attitudes that are consistently available nation-wide directly into travel demand models.

Several researchers have noted that similar households tend to cluster in homogeneous neighborhoods [2, 3, 11, 21], and that social networks influence work and residence location [3, 15, 42, 43]. Using the lifestyle cluster segmentation system of the targeted marketing data, we develop two variables to measure this tendency of households to choose to live near similar households and to live away from dissimilar households. We collapse the 26 household clusters in the lifestyle segmentation system into two variables, a similarity measure and a dissimilarity measure. The similarity measure is a standardized measure of the number of households in each choice alternative that share the same lifestyle cluster as the household making the residential location choice. The standardization accounts for differences in the alternatives' population sizes and differences in the relative sizes between each

cluster. Conversely, the dissimilarity measure is a standardized measure of the number of households in each alternative that are classified into the lifestyle cluster that is least correlated with that of the household making the choice. Technical details are in an appendix (see Section 5.8).

5.3.2 Mobile Phone Data

There are two major types of mobile phone data: (1) wireless signaling data and (2) global positioning system (GPS) data. In the case of wireless signaling data, location and time points are stored for all mobile phones not only when a phone call, text message, or email is sent or received, but also through a continual tracking process that allows service providers to more quickly connect phone calls. They do this by triangulating signals from cell towers, which is a process that creates less accurate location data than GPS data. The key benefit, though, is that data are collected for the whole population of mobile phone users regardless of the phones' specific capabilities or what applications are enabled. Additionally, location accuracy for places frequently visited like residence and workplace can be improved through repeat observations of a single phone over time. Cellular service providers and companies such as AirSage collect wireless signaling data; whereas travel time data providers like Google Maps, Inrix, and Waze primarily collect GPS data.

We use wireless signaling data in this study to calculate employment density for each location choice in the set of alternatives. For the 13-county Atlanta region, we obtained a home-work matrix at the grid cell level using latitude and longitude measurements. Each grid cell measures about 0.69 miles by 0.69 miles, or one-half of a square mile. We aggregated these grid cell counts by alternative and calculate the employment density at this level. We compared the latter variable with a traditional employment density variable from the metropolitan planning organization and determined no statistical difference in the significance of the variable or in model fit when substituting one for the other in the choice model.

Table 17: Mean Difference Between U.S. Census and Targeted Marketing Data

	Mean Difference
Income	
$0–$19,999	-3.52%
$20,000–$29,999	1.61%
$30,000–$39,999	2.36%
$40,000–$49,999	2.54%
$50,000–$74,999	4.00%
$75,000–$99,999	0.77%
$100,000–$149,999	-0.62%
$150,000+	-7.13%

5.3.3 U.S. Census Data

The U.S. Census data come from both the 2010 Decennial Census and the 2006-2010 American Community Survey (ACS) 5-year estimates. We use the 2010 Decennial Census to calculate the following neighborhood attributes: population density, the total number of housing units, and the average household size. We use the ACS data to estimate the average reported driving commute time for each alternative.

5.3.4 Representativeness of the Data

The final dataset contains 419,713 households, constituting a 25.0% sample of the 13-county metropolitan Atlanta region as compared to the 2010 Decennial Census, and 797 alternatives, which are delineated by U.S. Census tracts in the region. The targeted marketing data reflect characteristics of the Atlanta metropolitan area as of January 2011 and are fairly representative of the region. We show the mean difference between U.S. Census data and the targeted marketing data when comparing income groups at the tract level in Table 17. Additionally, we know from Kressner and Garrow [23] that the targeted marketing data do overrepresent homeowners, Whites, and individuals with at least some college education, but no more severely than the household travel survey does for the same region.

Although the sample is not fully representative of the population, it is still able to uncover relationships among variables in choice situations and quantify the effect of sample size on the sensitivity of choice models [1, 12]. It is not able to estimate the true share of

various homeowners or renters in the population. In particular, when the model is multinomial logit (MNL), Manski and Lerman [28] showed that the MNL parameter estimates obtained from a stratified sample will be consistent and unbiased relative to the MNL estimates obtained from a simple random sample under certain conditions. Thus, we do not expect that the model estimates will be impacted by the biases in our targeted marketing sample.

5.4 Methodology

In this study, we implement a multinomial logit (MNL) model. The utility V for household i in choosing alternative j from choice set J is a linear-in-parameters function of x_{ij}, $V_{ij} = x_{ij}\boldsymbol{\beta}+\epsilon$. If ϵ is assumed to be distributed independently and identically Gumbel (or extreme value type I), then the probability of individual i choosing alternative j is given as:

$$P_{ij} = \frac{e^{\boldsymbol{\beta}\mathbf{X}_{ij}}}{\sum_{j\epsilon J} e^{\boldsymbol{\beta}\mathbf{X}_{ij}}} \tag{9}$$

where $\boldsymbol{\beta}$ is a parameter vector and $\mathbf{X}_{ij}$ is a vector of variables for individual i and alternative j. The MNL model, although simple and elegant, requires the independence of irrelevant alternatives (IIA) property, which necessitates that eliminating an unchosen alternative will not affect the selection of the chosen alternative as the best option. The IIA property can be behaviorally unrealistic in many choice situations, particularly in a residential choice situation were spatial alternatives close to each other will likely have common unobserved spatial elements. A common specification for capturing such spatial correlation is to allow contiguous alternatives to be correlated [5]. In residential location choice models, the use of the MNL model is clearly not appropriate.

However, by invoking the IIA property, McFadden [29] showed that a random sample of alternatives without replacement including the chosen alternative, can be used to estimate an MNL model. Due to the high number of alternatives in spatial choice situations, the sampling of alternatives is appealing so that computation times are reduced. The MNL model's elegance and resulting ability to sample alternatives within the MNL framework has led to its continued use in the literature and its use here. Nerella and Bhat [31]

recommend that at a minimum one-eighth of the full choice set be used in the sampling of alternatives. Following this suggestion, we use a random sample of 100 alternatives in the base estimation. The following section describes how we proceed to test the sensitivity of this base model estimation to the number of random alternatives included in the estimation as well as the number of observations.

5.4.1 Sensitivity of Model Estimation

Nerella and Bhat [31] showed with the use of simulated data that parameter estimates can be effectively recovered with one-eighth of the full choice set. Because the data we use in the study provides a very large sample size, we have an opportunity to examine the sensitivity of the model estimation in a similar way. However, in addition to examining the effect of sample size of alternatives, we also examine the effect of the sample size of observations using real data.

Table 18 summarizes the sample size of observations used in recent literature for residential location choice modeling. All of the models estimated with more than 1,000 households shown in Table 18 come from three data sources: a National Household Travel Survey (NHTS) for the San Francisco Bay Area; the San Francisco Bay Area travel survey of 2000; or a Puget Sound Regional Council (PSRC) panel or activity survey for the Seattle, Washington region. To our knowledge, researchers generally build residential location choice models with small sample sizes in the range of 500 to 5,000 households.

In our study, both the number of observations and the number of alternatives in the estimation are systematically varied. The number of observations vary from 500 households to the full set of 419,713 households, and the number of sampled alternatives vary from 10 tracts to the full set of 797 tracts. For each combination of number of households and number of tracts, 10 random samples are drawn from the full set of households and tracts using a Monte Carlo method. The lower end of the scale of number of observations simulates the smaller sample sizes in recent literature. In total, we obtain parameter estimates and model statistics for 460 different datasets.

Table 18: Sample Sizes in Recent Residential Location Choice Literature

	Publication Year	Sample Size, N (Households)
Olaru, Smith, and Taplin [32]	2011	508
Ibeas, Cordera, DellOlio, and Coppola [18]	2013	534
Rashidi, Mohammadian, and Koppelman [37]	2011	615
Sener, Pendyala, and Bhat [39]	2011	702
Rashidi, Auld, and Mohammadian [38]	2012	741
Paleti, Bhat, and Pendyala [33]	2014	1,480[a]
Waddell, Bhat, Eluru, Wang, and Pendyala [47]	2007	1,823
Bhat, Paleti, Pendyala, Lorenzini, and Konduri [6]	2014	3,335
Lee and Waddell [26]	2010	4,739
Eluru, Bhat, Pendyala, and Konduri [9]	2010	5,082
Pinjari, Pendyala, Bhat, and Waddell [35]	2011	5,147
This study		419,713

[a] This model was estimated on employed individuals rather than households.

5.4.2 Model Assessment

As a measure of model fit, we use McFadden's likelihood ratio index, ρ, with respect to constants, defined as:

$$\rho = 1 - \frac{LL(\hat{\boldsymbol{\beta}})}{LL(0)} \tag{10}$$

where $LL(\hat{\boldsymbol{\beta}})$ is the value of the log-likelihood function at the estimated parameters and $LL(0)$ is its value when all the parameters are set equal to zero.

We utilize both the nested and non-nested log-likelihood ratio tests. The test statistic for the nested test is $-2(LL(\hat{\boldsymbol{\beta}}^H) - LL(\hat{\boldsymbol{\beta}}))$, where H indicates the model that is constrained or restricted according to the null hypothesis. If this test statistic exceeds the critical value of chi-squared with the appropriate degrees of freedom, the null hypothesis is rejected. In the case where a model cannot be written as a restricted version of the other that it is being tested against, Horowitz [17] proposed the following:

$$\Phi\left[-\Big(-2\left(\bar{\rho}_L - \bar{\rho}_S\right) \times LL(0) + \left(K_L - K_S\right)\Big)^{1/2}\right] \tag{11}$$

where Φ is the standard normal cumulative distribution function, $\bar{\rho}_L$ and $\bar{\rho}_S$ are the larger and smaller $\bar{\rho}$ values, respectively; and K_L and K_S are the number of parameters in the

model with the larger and smaller $\bar{\rho}$, respectively. McFadden's adjusted ρ, denoted $\bar{\rho}$, is defined as:

$$\bar{\rho} = 1 - \frac{LL(\hat{\boldsymbol{\beta}}) - K}{LL(0)} \tag{12}$$

where K is the number of parameters used in the model. If the value is less than the desired significance level, the null hypothesis is rejected.

5.5 Results

We present the results in two sections. First, we present the results for the basic MNL model specified in this study. Second, we present the results of the Monte Carlo experiment on the data, discussing the sensitivity of the model estimation to sample size of observations and number of sampled alternatives.

5.5.1 Basic MNL Model

Table 19 presents the results for three residential location choice models. Model 1 provides the simplest specification, and Models 2 and 3 add the similarity and dissimilarity measures.

The signs and significance of estimated parameters in these specifications are generally consistent with prior expectations. For the neighborhood attributes, population density is negative but not hugely significant, and the number of housing units is strongly positive. The sign on the neighborhood attribute of average income is negative due to the inclusion of the interaction of household income and average current market value. When this interaction is removed from the model, the sign on average income becomes positive.

For the accessibility variable, the estimated parameter is negative. In a traditional residential location choice model, the actual workplace is known for each household, and researchers calculate the commute time as a travel time between the known work location and the alternative. The commute variable in this case is much more influential in predicting a location choice because of this household specific information. In our case, we use the interaction of average commute time reported in the ACS for each Census tract and the employment density. This variable is likely only measuring a difference between urban and

Table 19: Summary of Model Results

	Model 1			Model 2			Model 3		
Variable	Parameter	Std Err	*t*-Stat	Parameter	Std Err	*t*-Stat	Parameter	Std Err	*t*-Stat
Neighborhood Attributes									
log(Average Income)	-0.213	0.004	-54.61	-0.079	0.004	-18.81	-0.125	0.004	-28.86
log(Population Density)	-0.047	0.002	-22.79	-0.073	0.002	-34.55	-0.067	0.002	-31.63
log(Number of Housing Units)	1.028	0.004	272.89	1.075	0.004	282.51	1.067	0.004	280.75
Accessibility Variables									
Avg Commute Time x *log*(Employment Density)	-0.042	0.006	-7.10	-0.159	0.006	-24.86	-0.182	0.006	-27.96
Interactions of Household and Alternative Characteristics									
Household Size x Avg Household Size	2.576	0.019	136.27	0.937	0.021	45.14	0.738	0.021	35.01
Household Income x Avg CMV[a]	0.318	0.001	227.86	0.126	0.002	78.73	0.122	0.002	76.23
Lifestyle Similarity Measure	—	—	—	0.574	0.001	471.90	0.474	0.001	332.38
Lifestyle Dissimilarity Measure	—	—	—	—	—	—	-0.388	0.003	-129.00
Model Statistics									
Number of Observations	419,713			419,713			419,713		
Number of Alternatives	100			100			100		
Log-likelihood with no Predictors	-1,932,850			-1,932,850			-1,932,850		
Log-likelihood at Convergence	-1,867,123			-1,771,893			-1,762,575		
Likelihood Ratio Index (ρ)	0.0340			0.0833			0.0881		

[a] CMV = current market value

suburban commute times and densities, which is why we anticipate that the significance is low.

The effect of the interaction of household size with the neighborhoods' average household size and the interaction of household income with the neighborhoods' average current market value are strongly positive in Model 1, with their effects lessening in Models 2 and 3. These interactions capture the tendency for clustering within neighborhoods by household size and income in Model 1. With the inclusion of the lifestyle similarity and dissimilarity measures in Models 2 and 3, we more directly account for this tendency.

To expand on this, a nested log-likelihood ratio test between Models 1 and 2 (with one degree of freedom) rejects the null hypothesis that the two models are equal at a 99.9% confidence level (-190,460 > (χ^2 = 10.8, df=1, 99.9%)), indicating that the similarity measure significantly improves the model fit. The similarity measure has a t-statistic of 471.90, which surpasses the significance of both the number of housing units available (t-stat = 282.51) and the interaction of household income with the neighborhoods' average current market value (t-stat = 78.73) in Model 1. The parameter estimate is positive, correctly reflecting that having a high occurrence of similar households in a choice alternative is a desirable attribute.

Furthermore, the dissimilarity measure added in Model 3 is negative and significant, indicting that having a high occurrence of dissimilar households in a choice location is an undesirable attribute. This measure impacts the choice of location less than the similarity measure. By including the dissimilarity measure, the significance is ultimately split between the similarity and dissimilarity measures, but the model fit still improves. A nested log-likelihood ratio test between Models 2 and 3 rejects the null hypothesis that they are equal at a 99.9% confidence level (18,636 > (χ^2 = 10.8, df=1, 99.9%)). The significant relative increases in ρ of Models 2 and 3 over Model 1 and the high significance of the variables themselves demonstrates that the lifestyle variables help in understanding residential location behavior.

5.5.2 Monte Carlo Experiment

We use Model 3 from Table 19 in the Monte Carlo experiment to test the variance of model estimation due to sample sizes of the random alternatives included in the estimate as well as the number of observations. For each combination of sample sizes, we compile ten different datasets, estimate the model specification from Model 3, and summarize the results. Tables 34, 30, and 36 present representative results for the Monte Carlo simulations for the parameter estimates. In Tables 34 and 30, the variables with the largest variation in the standard deviation of the estimates across all of the runs is summarized; whereas in Table 36, the variable with the smallest variation is summarized. The full set of simulation results can be found in Appendices A and B. Additionally, Table 23 summarizes how ρ varies across the models, and Table 24 summarizes the computation times across the model estimations.

In general, as the number of observations increase, the standard deviation of the parameter estimates decreases, as expected. Importantly, the effect of the number of observations on the variability of the estimates is significantly greater than the effect of the number of sampled alternatives. In Table 34, the parameter estimates for the interaction of household size and average household size for each alternative are summarized for the experiment. The top left quadrant shows the mean of all the model runs by sample sizes of observations and alternatives. The lower left quadrant shows the standard deviations, the upper right quadrant shows the minimum, and the lower right quadrant shows the maximum of all the runs. Let us assume from the mean of the 419,713 model runs that the "true" parameter estimate is near 0.73. Relative to this mean, the standard deviations seen in the smaller household sample sizes (which match those generally seen in recent literature) were high. Increasing the number of sampled alternatives did not consistently lower the standard deviation. However, increasing the number of sampled households did, and we saw large decreases in standard deviations with these larger household samples.

Similarly, the average income shown in Table 30 had such a high standard deviation relative to its "true" value that for the smaller household sample sizes, we saw the sign on the estimate switch. In other words, given a particular random sample of just 500 or 1,000

Table 20: Household Size x Avg Household Size

	Alts						Alts					
Obs	10	50	100	200	400	797	10	50	100	200	400	797
	Mean of Estimates						Minimum of Estimates					
500	1.131	0.724	0.474	0.569	0.930	0.913	0.120	0.454	-0.058	-0.020	0.260	0.094
1,000	0.330	0.802	0.709	0.629	0.705	0.856	-0.346	0.301	0.163	0.193	0.079	0.161
2,500	0.724	0.734	0.714	0.758	0.819	0.681	0.379	0.508	0.298	0.333	0.663	0.215
5,000	0.700	0.793	0.718	0.694	0.790	0.748	0.520	0.594	0.480	0.575	0.409	0.476
10,000	0.640	0.699	0.676	0.729	0.783	0.791	0.468	0.543	0.549	0.520	0.660	0.558
50,000	0.729	0.753	0.769	0.741	0.767	0.795	0.673	0.645	0.706	0.641	0.686	0.729
100,000	0.724	0.735	0.736	0.753	—	—	0.645	0.695	0.706	0.682	—	—
200,000	0.723	0.739	0.727	—	—	—	0.679	0.699	0.686	—	—	—
419,713	0.728	0.730	0.731	—	—	—	0.719	0.721	0.725	—	—	—
	Standard Deviation of Estimates						Maximum of Estimates					
500	0.526	0.196	0.400	0.528	0.410	0.386	1.701	1.011	1.166	1.611	1.366	1.348
1,000	0.495	0.408	0.460	0.285	0.405	0.438	1.345	1.581	1.658	1.091	1.318	1.546
2,500	0.213	0.133	0.207	0.234	0.140	0.315	1.101	0.956	1.082	1.065	1.126	1.146
5,000	0.120	0.152	0.161	0.082	0.256	0.180	0.915	1.027	1.039	0.828	1.183	0.972
10,000	0.117	0.116	0.104	0.108	0.095	0.130	0.786	0.899	0.861	0.882	0.932	0.968
50,000	0.054	0.075	0.036	0.058	0.053	0.055	0.810	0.857	0.807	0.836	0.857	0.924
100,000	0.042	0.033	0.015	0.035	—	—	0.788	0.782	0.762	0.803	—	—
200,000	0.032	0.032	0.021	—	—	—	0.779	0.795	0.762	—	—	—
419,713	0.008	0.004	0.003	—	—	—	0.739	0.736	0.736	—	—	—

— = Due to limitations of the random access memory (RAM) on the computer used for estimation, these models were not estimated.

Table 21: *log*(Average Income)

	Alts						Alts					
Obs	10	50	100	200	400	797	10	50	100	200	400	797
	Mean of Estimates						Minimum of Estimates					
500	-0.272	-0.145	-0.101	-0.087	-0.109	-0.123	-0.575	-0.236	-0.304	-0.221	-0.291	-0.347
1,000	-0.144	-0.128	-0.092	-0.099	-0.113	-0.177	-0.280	-0.234	-0.214	-0.221	-0.280	-0.300
2,500	-0.135	-0.129	-0.125	-0.120	-0.114	-0.118	-0.225	-0.274	-0.180	-0.176	-0.179	-0.147
5,000	-0.179	-0.121	-0.121	-0.125	-0.111	-0.122	-0.239	-0.172	-0.169	-0.183	-0.162	-0.216
10,000	-0.161	-0.123	-0.129	-0.118	-0.118	-0.130	-0.184	-0.156	-0.155	-0.179	-0.140	-0.206
50,000	-0.154	-0.126	-0.116	-0.125	-0.128	-0.136	-0.170	-0.143	-0.132	-0.141	-0.145	-0.153
100,000	-0.156	-0.133	-0.124	-0.126	—	—	-0.166	-0.142	-0.134	-0.133	—	—
200,000	-0.151	-0.132	-0.125	—	—	—	-0.160	-0.135	-0.129	—	—	—
419,713	-0.156	-0.131	-0.125	—	—	—	-0.159	-0.132	-0.126	—	—	—
	Standard Deviation of Estimates						Maximum of Estimates					
500	0.143	0.067	0.138	0.111	0.127	0.135	-0.062	-0.009	0.089	0.102	0.126	0.049
1,000	0.123	0.078	0.085	0.088	0.098	0.080	0.124	-0.007	0.087	0.031	0.012	-0.046
2,500	0.051	0.074	0.049	0.045	0.039	0.023	-0.052	-0.034	-0.052	-0.039	-0.074	-0.080
5,000	0.042	0.040	0.035	0.032	0.030	0.042	-0.116	-0.041	-0.059	-0.090	-0.074	-0.081
10,000	0.016	0.021	0.018	0.031	0.014	0.032	-0.138	-0.095	-0.101	-0.078	-0.089	-0.101
50,000	0.009	0.017	0.010	0.011	0.009	0.012	-0.143	-0.092	-0.102	-0.108	-0.118	-0.115
100,000	0.007	0.007	0.006	0.004	—	—	-0.141	-0.123	-0.115	-0.118	—	—
200,000	0.005	0.004	0.002	—	—	—	-0.146	-0.124	-0.121	—	—	—
419,713	0.002	0.001	0.001	—	—	—	-0.154	-0.129	-0.124	—	—	—

— = Due to limitations of the random access memory (RAM) on the computer used for estimation, these models were not estimated.

Table 22: Lifestyle Similarity Measure

	Alts						Alts					
Obs	10	50	100	200	400	797	10	50	100	200	400	797
	Mean of Estimates						Minimum of Estimates					
500	0.559	0.489	0.491	0.467	0.436	0.428	0.502	0.452	0.443	0.421	0.378	0.380
1,000	0.559	0.489	0.484	0.459	0.449	0.407	0.510	0.439	0.443	0.430	0.414	0.372
2,500	0.533	0.483	0.473	0.463	0.442	0.431	0.492	0.463	0.442	0.450	0.415	0.405
5,000	0.541	0.485	0.473	0.459	0.452	0.435	0.499	0.473	0.436	0.441	0.433	0.405
10,000	0.543	0.490	0.476	0.459	0.452	0.434	0.529	0.474	0.468	0.438	0.445	0.416
50,000	0.541	0.487	0.471	0.462	0.449	0.433	0.534	0.480	0.469	0.453	0.444	0.425
100,000	0.543	0.488	0.472	0.463	—	—	0.540	0.486	0.469	0.460	—	—
200,000	0.543	0.487	0.473	—	—	—	0.539	0.483	0.472	—	—	—
419,713	0.542	0.487	0.473	—	—	—	0.541	0.486	0.472	—	—	—
	Standard Deviation of Estimates						Maximum of Estimates					
500	0.054	0.035	0.039	0.030	0.028	0.030	0.639	0.556	0.561	0.502	0.469	0.485
1,000	0.037	0.028	0.029	0.014	0.025	0.025	0.610	0.515	0.548	0.480	0.485	0.442
2,500	0.024	0.018	0.019	0.010	0.015	0.016	0.559	0.514	0.496	0.477	0.462	0.451
5,000	0.017	0.013	0.017	0.013	0.013	0.015	0.560	0.511	0.488	0.487	0.473	0.460
10,000	0.010	0.010	0.006	0.011	0.006	0.012	0.562	0.506	0.489	0.473	0.463	0.449
50,000	0.005	0.004	0.001	0.005	0.002	0.004	0.550	0.494	0.473	0.468	0.452	0.439
100,000	0.002	0.002	0.002	0.002	—	—	0.546	0.492	0.476	0.465	—	—
200,000	0.002	0.002	0.001	—	—	—	0.547	0.490	0.476	—	—	—
419,713	0.001	0.001	0.001	—	—	—	0.544	0.488	0.474	—	—	—

— = Due to limitations of the random access memory (RAM) on the computer used for estimation, these models were not estimated.

Table 23: Likelihood Ratio Index (ρ)

	Alts						Alts					
Obs	10	50	100	200	400	797	10	50	100	200	400	797
	Mean of Estimates						Minimum of Estimates					
500	0.1658	0.1051	0.0914	0.0745	0.0688	0.0606	0.1364	0.0861	0.0799	0.0619	0.0621	0.0507
1,000	0.1610	0.1024	0.0886	0.0778	0.0682	0.0595	0.1487	0.0966	0.0827	0.0719	0.0602	0.0534
2,500	0.1576	0.1015	0.0856	0.0772	0.0682	0.0601	0.1510	0.0944	0.0805	0.0731	0.0641	0.0565
5,000	0.1574	0.1026	0.0895	0.0766	0.0673	0.0607	0.1492	0.0981	0.0776	0.0742	0.0641	0.0585
10,000	0.1595	0.1032	0.0891	0.0759	0.0685	0.0602	0.1540	0.1004	0.0858	0.0725	0.0668	0.0583
50,000	0.1605	0.1027	0.0877	0.0769	0.0677	0.0603	0.1578	0.1013	0.0866	0.0758	0.0668	0.0588
100,000	0.1595	0.1030	0.0880	0.0769	—	—	0.1577	0.1014	0.0874	0.0758	—	—
200,000	0.1600	0.1030	0.0880	—	—	—	0.1595	0.1025	0.0878	—	—	—
419,713	0.1596	0.1028	0.0881	—	—	—	0.1591	0.1027	0.0880	—	—	—
	Standard Deviation of Estimates						Maximum of Estimates					
500	0.0154	0.0109	0.0063	0.0052	0.0051	0.0062	0.1797	0.1251	0.0975	0.0803	0.0765	0.0708
1,000	0.0110	0.0036	0.0042	0.0035	0.0052	0.0038	0.1767	0.1072	0.0947	0.0830	0.0785	0.0656
2,500	0.0053	0.0042	0.0042	0.0022	0.0028	0.0029	0.1664	0.1070	0.0928	0.0804	0.0724	0.0653
5,000	0.0046	0.0030	0.0053	0.0017	0.0024	0.0015	0.1628	0.1063	0.0982	0.0790	0.0709	0.0627
10,000	0.0032	0.0023	0.0018	0.0019	0.0012	0.0010	0.1638	0.1085	0.0912	0.0788	0.0706	0.0617
50,000	0.0014	0.0008	0.0007	0.0006	0.0005	0.0007	0.1621	0.1037	0.0888	0.0776	0.0685	0.0611
100,000	0.0015	0.0009	0.0004	0.0005	—	—	0.1621	0.1042	0.0885	0.0777	—	—
200,000	0.0005	0.0004	0.0002	—	—	—	0.1610	0.1036	0.0883	—	—	—
419,713	0.0003	0.0001	0.0000	—	—	—	0.1601	0.1029	0.0881	—	—	—

— = Due to limitations of the random access memory (RAM) on the computer used for estimation, these models were not estimated.

Table 24: Computation Time (Mins)

	Alts					
Obs	10	50	100	200	400	797
Mean						
500	0.00	0.01	0.03	0.07	0.18	0.59
1,000	0.00	0.02	0.05	0.12	0.35	1.22
2,500	0.01	0.06	0.13	0.32	0.98	3.26
5,000	0.02	0.10	0.25	0.66	1.98	6.65
10,000	0.03	0.21	0.49	1.30	3.94	13.26
50,000	0.17	1.07	2.54	6.91	20.72	379.17
100,000	0.33	2.23	5.23	13.76	—	—
200,000	0.72	4.44	10.50	—	—	—
419,713	1.51	9.46	194.50	—	—	—
Standard Deviation						
500	0.03	0.05	0.07	0.18	0.42	0.94
1,000	0.06	0.05	0.09	0.22	0.70	2.05
2,500	0.19	0.11	0.35	0.74	0.40	6.17
5,000	0.35	0.24	0.40	0.39	1.02	2.19
10,000	0.53	0.47	0.43	2.62	2.13	5.17
50,000	1.40	1.17	2.68	5.16	10.40	1875.90
100,000	4.13	3.17	3.68	12.38	—	—
200,000	6.99	4.39	9.24	—	—	—
419,713	9.08	6.76	1981.81	—	—	—

— = Due to limitations of the random access memory (RAM) on the computer used for estimation, these models were not estimated.

households, it is likely that the parameter estimate will have the opposite sign of the true value. In this case, the estimate no longer contributed a disutility and actually contributed a positive utility.

In the case of the last example shown in Table 36, the results were the same but less severely. The standard deviation was relatively consistent across the number of sampled alternatives, and decreased as the number of sampled households increased. The full set of 797 alternatives with a sample of 500 households still produced a relatively high standard deviation (0.030) as compared to the estimates from 10 alternatives and the full set of 719,713 households (0.001).

Lastly, the estimates appear to stabilize around 50,000 or 100,000 observations in this particular study, but this determination is dependent on the tolerance of the particular application and likely the specification. The increase in computation time for estimating the model with 100,000 households rather than 500 (with 100 sample alternatives) only increases by an average of 5.2 minutes.

5.6 Conclusions

This study contributes to the literature in several ways. First, it is among the initial few studies in the field of transportation that utilizes targeted marketing data in transportation planning applications. We expect that the integration of targeted marketing data with other types of data will allow researchers to investigate questions related to travel behavior that are not currently possible with traditional household travel surveys.

Second, this study demonstrates that targeted marketing data can provide new behavioral insights through a large number of variables that significantly improve our understanding of residential location behavior. This study, supported by results from our other application in travel demand modeling [22], suggests that the new data might improve understanding of many travel behaviors researchers model in transportation planning.

Third, this study emphasizes the importance of large sample sizes in specifying robust models. The effect of sample size on model variance is so large that discussions relating to the number of alternatives to sample is almost peripheral. Targeted marketing data provide

the opportunity to obtain large sample sizes of detailed household-level data with coverage rates at an estimated 87.5% as compared to 2010 Decennial Census data at low costs [23].[1] Compared to traditional household travel surveys with sample sizes in the range of 1% or less, inexpensive targeted marketing data is worth exploring further.

Future research needs to address the limitations of targeted marketing data. The use of targeted marketing data and combinations of it with other traditional or third-party data will be relatively straightforward for some applications, such as the one examined in this work, but will be more challenging for others. Many modeling applications require more information than is available in targeted marketing data. Most obviously, targeted marketing data lacks real trip-making behavior. Future research must identify techniques for overcoming the lack of trip-making behavior— such as building populations that combine the sociodemographic information available in targeted marketing data with observed trip patterns available through traditional household travel surveys and/or GPS and other tracking devices.

5.7 Acknowledgements

The authors would like to express their deep appreciation to AirSage Inc., the provider of the mobile phone data, and to the targeted marketing firm that has been helpful in preparing its data for our use. This work was partially supported by a National Science Foundation Graduate Research Fellowship (NSF GRFP).

5.8 Appendix

5.8.1 Calculation of Similarity Measure

To measure the tendency of households to move towards those that are most like them, a "similarity measure," is created. The following procedure explains each step in the calculation of the measure. Table 25 shows simplified household-level data, which will be used as an example to demonstrate the calculation. Ten households, four geographies, and three

[1] We define a coverage rate as the sample size over the total estimate of persons or households by the Decennial Census, expressed at a percentage. This definition conforms to that of the [46]. For clarity, a coverage rate of 87.5% means that 87.5% of people in the Decennial Census are present in the targeted marketing data.

Table 25: Simplified Household-Level Data for Example Calculations

Household	Geography	Cluster
01	A	2
02	B	1
03	B	2
04	B	2
05	C	2
06	C	2
07	C	2
08	C	3
09	D	1
10	D	2

clusters are given in this simplified example. In actuality, 419,713 households, 797 Census tracts, and 26 lifestyle clusters are used.

1. A cross tabulation of geography by cluster is calculated (Table 26a).

2. The cross tabulation is standardized in two steps.

    a. Each element is divided by its row sum to control for differences in area and population between geographies (Table 26b).

    b. A z-score is calculated for each element, i, by column, j, which is given by:

$$z_{i,j} = \frac{x_{i,j} - \mu_j}{\sigma_j}, \tag{13}$$

    where μ_j is the mean of column j and σ_j is its standard deviation. This controls for differences in the number of households classified into each cluster (Table 26c). For the element in geography A, cluster 1:

$$\mu_1 = 0.2075$$

$$\sigma_1 = 0.2494$$

$$z_{A,1} = \frac{0.00 - 0.2075}{0.2494} = -0.83$$

This table is subsequently used when populating the alternative-specific similarity variable. For a given household classified in cluster X, a z-score for cluster X is obtained from the

Table 26: Example Calculation of Double-Standardized Cross Tabulation

	Cluster		
Geography	1	2	3
A	0	1	0
B	1	2	0
C	0	3	1
D	1	1	0

(a) Step 1

	Cluster		
Geography	1	2	3
A	0.00	1.00	0.00
B	0.33	0.67	0.00
C	0.00	0.75	0.25
D	0.50	0.50	0.00

(b) Step 2a

	Cluster		
Geography	1	2	3
A	-0.83	1.30	-0.50
B	0.50	-0.30	-0.50
C	-0.83	0.10	1.50
D	1.17	-1.10	-0.50

(c) Step 2b

table for each geography in their choice set. The results can be seen in Table 28 for the example.

5.8.2 Calculation of Dissimilarity Measure

To measure the tendency of households to choose residential neighborhoods with a small number of households that are unlike their own, a "dissimilarity measure" is created. Just like the similarity measure, this variable collapses the 26 households clusters into one variable. This measure uses the same standardized table calculated in Section 5.8.1 (Table 26c). However, rather than taking the z-score for the lifestyle cluster of the household under consideration, the z-score is taken for the lifestyle cluster that is least-correlated with their lifestyle cluster. The following procedure demonstrates the process for finding the least-correlated cluster, or most dissimilar, for the example scenario.

1. A correlation matrix of the clusters is calculated (Table 27a).

2. The least-correlated cluster is selected for each cluster. This is equivalent to the smallest number in each row of the matrix from Step 1 (Table 27b).

Table 27c shows the related table for the 26 lifestyle clusters in the targeted marketing data. This table correlates with the example data in Table 27b. Table 28 shows the dissimilarity variable for the example. For a given household classified in cluster X, the most dissimilar cluster is identified, cluster Y. A z-score for cluster Y is obtained from the table for each geography in their choice set.

Table 27: Most Dissimilar Clusters

Cluster	Cluster 1	2	3
1	1.00	-0.87	-0.56
2	-0.87	1.00	0.07
3	-0.56	0.07	1.00

(a) Step 1

Cluster	Most Dissimilar
1	2
2	1
3	1

(b) Step 2

Cluster	Most Dissimilar
Already Affluent	Mid-Life Munchkins
Big Spender Parents	Rocky Road
Chic Society	Very Spartan
Diamonds-to-Go	Very Spartan
Easy Street	Rocky Road
Feathering-the-Nest	Very Spartan
Go-Go Families	Quiet Homebodies
Home Hoppers	Mid-Life Munchkins
IRA Spenders	Very Spartan
Just Sailing Along	Mid-Life Munchkins
Kiddie Kastles	Rocky Road
Loose Change	Still Going Strong
Mid-Life Munchkins	Rocky Road

Cluster	Most Dissimilar
Nice & Easy Grandparents	Very Spartan
Oodles of Offspring	Diamonds-to-Go
Parks, Parts, & Prayers	Diamonds-to-Go
Quiet Homebodies	Totebaggers
Rocky Road	Kiddie Kastles
Still Going Strong	Loose Change
Totebaggers	Kiddie Kastles
Under-the-Car	Diamonds-to-Go
Very Spartan	Diamonds-to-Go
Working Hard	Diamonds-to-Go
X-tra Needy	Diamonds-to-Go
Young-at-Heart	Diamonds-to-Go
Zero Mobility	Diamonds-to-Go

(c) Step 2 for lifestyle clusters in targeted marketing data

Table 28: "Similarity Measure" and "Dissimilarity Measure" for Example Calculation

		Similarity Measure by Geography					Dissimilarity Measure by Geography			
Household	Cluster	A	B	C	D	Dissimilar	A	B	C	D
01	2	1.30	-0.30	0.10	-1.10	1	-0.83	0.50	-0.83	1.17
02	1	-0.83	0.50	-0.83	1.17	2	1.30	-0.30	0.10	-1.10
03	2	1.30	-0.30	0.10	-1.10	1	-0.83	0.50	-0.83	1.17
04	2	1.30	-0.30	0.10	-1.10	1	-0.83	0.50	-0.83	1.17
05	2	1.30	-0.30	0.10	-1.10	1	-0.83	0.50	-0.83	1.17
06	2	1.30	-0.30	0.10	-1.10	1	-0.83	0.50	-0.83	1.17
07	2	1.30	-0.30	0.10	-1.10	1	-0.83	0.50	-0.83	1.17
08	3	-0.50	-0.50	1.50	-0.50	1	-0.83	0.50	-0.83	1.17
09	1	-0.83	0.50	-0.83	1.17	2	1.30	-0.30	0.10	-1.10
10	2	1.30	-0.30	0.10	-1.10	1	-0.83	0.50	-0.83	1.17

A weighted average dissimilarity variable, which takes into account the z-score for a subset of the least-correlated clusters, was also tested. The use of one to five of the most dissimilar clusters was tested. The resulting five models were not statistically different from one another. A non-nested log-likelihood ratio test between the model with a single-cluster dissimilarity measure and the model with a five-cluster weighted average dissimilarity measure, which was the model with the largest log-likelihood at convergence value, failed to reject the null hypothesis that they were equal ($4.4 > (\Phi{=}3.1, 99.9\%)$). Therefore, it was determined that adding additional dissimilar clusters to the calculation of the dissimilarity variable did not improve the model fit enough to outweigh the additional computation time.

5.9 References

[1] E. R. Babbie. *The Practice of Social Research*. Cengage Learning, 13th edition, 2012.

[2] P. Bayer and R. McMillan. "Tiebout sorting and neighborhood stratification." *Journal of Public Economics*, 96:1129–1143, 2012.

[3] P. Bayer, S. L. Ross, and G. Topa. "Place of Work and Place of Residence: Informal Hiring Networks and Labor Market Outcomes." *Journal of Political Economy*, 116(6): 1150–1196, 2008.

[4] C. R. Bhat. "The Maximum Approximate Composite Marginal Likelihood (MACML) Estimation of Multinomial Probit-Based Unordered Response Choice Models." *Transportation Research Part B: Methodological*, 45(7):923–939, 2011.

[5] C. R. Bhat and J. Guo. "A mixed spatially correlated logit model: formulation and application to residential choice modeling." *Transportation Research Part B: Methodological*, 38:147–168, 2004.

[6] C. R. Bhat, R. Paleti, R. M. Pendyala, K. Lorenzini, and K. C. Konduri. "Accommodating Immigration Status and Self Selection Effects in a Joint Model of Household Auto Ownership and Residential Location Choice." *Transportation Research Record: Journal of the Transportation Research Board*, 2014. Forthcoming.

[7] S. Binder, G. S. Macfarlane, L. A. Garrow, and M. Bierlaire. "Associations Among Household Characteristics, Vehicle Characteristics, and Emission Failures: An Application of Targeted Marketing Data." *Transportation Research Part A: Policy and Practice*, 59:122–133, 2014.

[8] DELTA. "DELTA applications." `http://www.davidsimmonds.com/`, 2014. Accessed on 27 January 2014.

[9] N. Eluru, C. R. Bhat, R. M. Pendyala, and K. C. Konduri. "A joint flexible econometric model system of household residential location and vehicle fleet composition/usage choices." *Transportation*, 37:603–626, 2010.

[10] Federal Highway Administration. "2009 National Household Travel Survey (NHTS) User's Guide (Version 2)." http://nhts.ornl.gov/2009/pub/UsersGuideV2.pdf, 2011. Accessed on 3 September 2013.

[11] A. Frenkel, E. Bendit, and S. Kaplan. "The linkage between the lifestyle of knowledge-workers and their intra-metropolitan residential choice: A clustering approach based on self-organizing maps." *Computers, Environment and Urban System*, 39:151–161, 2013.

[12] R. M. Groves. *Survey Errors and Survey Costs*. Wiley-Interscience, 2004.

[13] C. A. Guevara and M. Ben-Akiva. "Endogeneity in Residential Location Choice Models." *Transportation Research Record: Journal of the Transportation Research Board*, 1977:60–66, 2006.

[14] C. A. Guevara and M. Ben-Akiva. "Sampling of alternatives in Multivariate Extreme Value (MEV) models." *Transportation Research Part B: Methodological*, 48: 31–52, 2013.

[15] J. Hackney and K. Axhausen. "An agent model of social network and travel behavior interdependence." In *11th International Conference of the International Association for Travel Behavior Research (IATBR)*, 2006.

[16] HBA Specto Incorporated. "Land Use, Transport and Spatial Economic Modelling." http://www.hbaspecto.com/pecas/, 2014. Accessed on 27 January 2014.

[17] J. Horowitz. "An evaluation of the usefulness of two standard goodness-of-fit indicators for comparing non-nested random utility models." *Transportation Research Record: Journal of the Transportation Research Board*, 874:19–25, 1982.

[18] Á. Ibeas, R. Cordera, L. DellOlio, and P. Coppola. "Modeling the spatial interactions between workplace and residential location." *Transportation Research Part A: Policy and Practice*, 49:110–122, 2013.

[19] International Telecommunication Union (ITU). "Time Series by Country: Fixed-Telephone Subscriptions 2000-2012." http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx, 2013. Accessed on 3 September 2013.

[20] International Telecommunication Union (ITU). "Time Series by Country: Mobile-Cellular Subscriptions 2000-2012." http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx, 2014. Accessed on 21 January 2014.

[21] H. Kiefer. "Residential Location Choice: The Role of a Taste for Similarity." *International Journal of Economics and Finance*, 4(9), 2012.

[22] J. D. Kressner and L. A. Garrow. "Lifestyle Segmentation Variables as Predictors of Home-Based Trips for Atlanta, Georgia Airport." *Transportation Research Record: Journal of the Transportation Research Board*, 2266:20–30, 2012.

[23] J. D. Kressner and L. A. Garrow. "Using Third-Party Data for Travel Demand Modeling: A Comparison of Targeted Marketing, Census, and Household Travel Survey Data." *Transportation Research Record: Journal of the Transportation Research Board*, 2014. Accepted.

[24] J. D. Kressner and L. A. Garrow. "Leveraging targeted marketing data in travel demand modeling: An application in residential location choice modeling.". Georgia Institute of Technology. Working paper, 2014.

[25] J. D. Kressner, M. F. Carragher, and K. E. Watkins. "A Household-Level Pairwise Comparison of Targeted Marketing Data and Self-Reported Survey Data." In *Proceedings of the 2014 Annual Meeting of the Transportation Research Board*, 2014.

[26] B. H. Y. Lee and P. Waddell. "Residential mobility and location choice: a nested logit model with sampling of alternatives." *Transportation*, 37:587–601, 2010.

[27] J. D. Lemp and K. M. Kockelman. "Strategic sampling for large choice sets in estimation and application." *Transportation Research Part A: Policy and Practice*, 46: 602–613, 2012.

[28] C. F. Manski and S. R. Lerman. "The Estimation of Choice Probabilities from Choice Based Samples." *Econometrica*, 45(8):1977–1988, 1977.

[29] D. McFadden. "Modeling the choice of residential location." In A. Karlqvist, L. Lundqvist, F. Snickers, and J. W. Weibull, editors, *Spatial Interaction Theory and Planning Models*, pages 75–96. North-Holland Publishing Company, 1978.

[30] P. L. Mokhtarian and X. Cao. "Examining the impacts of residential self-selection on travel behavior: A focus on methodologies." *Transportation Research Part B: Methodological*, 42:204–228, 2008.

[31] S. Nerella and C. R. Bhat. "Numerical Analysis of Effect of Sampling of Alternatives in Discrete Choice Models." *Transportation Research Record: Journal of the Transportation Research Board*, 1894:11–19, 2004.

[32] D. Olaru, B. Smith, and J. H. Taplin. "Residential location and transit-oriented development in a new rail corridor." *Transportation Research Part A: Policy and Practice*, 45:219–237, 2011.

[33] R. Paleti, C. R. Bhat, and R. M. Pendyala. "An Integrated Model of Residential Location, Work Location, Vehicle Ownership, and Commute Tour Characteristics." *Transportation Research Record: Journal of the Transportation Research Board*, 2014. Forthcoming.

[34] Pew Research Center. "Cell phone ownership hits 91% of adults." `http://www.pewresearch.org/fact-tank/2013/06/06/cell-phone-ownership-hits-91-of-adults/`, 2014. Accessed on 21 January 2014.

[35] A. R. Pinjari, R. M. Pendyala, C. R. Bhat, and P. A. Waddell. "Modeling the choice continuum: an integrated model of residential location, auto ownership, bicycle ownership, and commute tour mode choice decisions." *Transportation*, 38:933–958, 2011.

[36] S. Putman. "METROPolitan Integrated Land Use System." `http://gis.kent.edu/gis/empact/lit_urb_md01.htm`, 2014. Accessed on 27 January 2014.

[37] T. H. Rashidi, A. Mohammadian, and F. S. Koppelman. "Modeling interdependencies between vehicle transaction, residential relocation and job change." *Transportation*, 38: 909–932, 2011.

[38] T. H. Rashidi, J. Auld, and A. Mohammadian. "A behavioral housing search model: Two-stage hazard-based and multinomial logit approach to choice-set formation and location selection." *Transportation Research Part A: Policy and Practice*, 46:1097–1107, 2012.

[39] I. N. Sener, R. M. Pendyala, and C. R. Bhat. "Accommodating spatial correlation across choice alternatives in discrete choice models: an application to modeling residential location choice behavior." *Journal of Transport Geography*, 19:294–303, 2011.

[40] P. Stopher. *Collecting, Managing, and Assessing Data Using Sample Surveys*. Cambridge University Press, 2012.

[41] P. R. Stopher and S. P. Greaves. "Household travel surveys: Where are we going?." *Transportation Research Part A: Policy and Practice*, 41:367–381, 2007.

[42] H. T. Theo Arentze. "Social networks, social interactions, and activity-travel behavior: a framework for microsimulation." *Environment and Planning B: Planning and Design*, 35(6):1012–1027, 2008.

[43] N. Tilahun and D. Levinson. "Work and home location: Possible role of social networks." *Transportation Research Part A: Policy and Practice*, 45:323–331, 2011.

[44] Travel Survey Methods Committee (ABJ40). "The Online Travel Survey Manual: A Dynamic Document for Transportation Professionals." `http://www.travelsurveymanual.org/`, 2014. Accessed on 20 January 2014.

[45] UrbanSim. "UrbanSim Overview." `http://www.urbansim.org/Main/UrbanSim`, 2014. Accessed on 27 January 2014.

[46] U.S. Census Bureau. "American Community Survey Coverage Rates: Definitions." `http://www.census.gov/acs/www/methodology/coverage_rates_definitions/`, 2013. Accessed on 14 November 2013.

[47] P. Waddell, C. R. Bhat, N. Eluru, L. Wang, and R. M. Pendyala. "Modeling interdependence in household residence and workplace choices." *Transportation Research Record: Journal of the Transportation Research Board*, 2003:84–92, 2007.

[48] J. Zmud, M. Lee-Gosselin, J. A. Carrasco, and M. A. Munizaga, editors. *Transport Survey Methods, Best Practice for Decision Making*, 2013. Proceedings from the 9th International Conference on Transport Survey Methods.

CHAPTER VI

CONCLUSION

6.1 Review

This work accomplished two main research objectives: to validate targeted marketing data by studying its representativeness of population characteristics and to test the usability and effectiveness of targeted marketing data in travel demand modeling applications. Chapters 2 and 3 focused on the first objective and Chapters 4 and 5 focused on the second objective. Chapter 2 completed an aggregate comparison of targeted marketing data to U.S. Census data and a household travel survey. Chapter 3 reported on the pairwise household-level comparison between targeted marketing data and self-reported survey data for a population group that is historically hard-to-reach. Chapter 4 investigated the use of targeted marketing data, with the associated lifestyle clusters, in an airport passenger model. Chapter 5 investigated the use of targeted marketing data in a basic residential location choice model. It also tested model sensitivity with varying numbers of observations and alternatives included in the model using Monte Carlo simulations. This chapter summarizes the major findings of each of these studies and outlines directions for future research.

6.2 Major Conclusions and Future Research

6.2.1 Aggregate Validation

The results from this study showed that the distributions of demographic and socioeconomic variables are similar between targeted marketing and U.S. Census data, particularly for age, gender, household income, and the presence of children. The largest discrepancies are associated with educational attainment and ethnicity, which are variables rarely used in traditional transportation planning applications. The higher discrepancies for those in the middle categories of educational attainment as compared to those at the far ends suggests that these discrepancies are associated with imputation rates, and moreover that the imputation methods do not rely primarily on Census data. The targeted marketing

data do contain fewer individuals under the age of 40, fewer low income households, and fewer households with children than U.S. Census data, but these discrepancies are small. For all of the variables, the median differences observed at the block group or tract levels are all within 12.5%, with the largest differences arising in housing type (tenure).

Due largely to the smaller sample sizes of the household travel survey, the deviation of differences over the small aggregation areas is higher for the household travel survey comparison than the mirroring targeted marketing data comparison in all cases except for educational attainment. Similar to the targeted marketing data findings, the household travel survey data also contain fewer individuals under the age of 40, fewer low-income households, and fewer households with children than the U.S. Census data. Interestingly, the targeted marketing and household travel survey data both underrepresent renters and overrepresent Whites to a very similar degree.

The weighted household travel survey data matches Census data somewhat more closely than the unweighted household travel survey data. Because the sample size of the household travel survey was so small, the weighting has little effect on the overall distribution of the differences. However, if a similar weighting technique was adopted for the targeted marketing data, the biases present in the targeted marketing data could be significantly reduced. Future research should investigate this hypothesis.

Overall, the distributions of demographic and socioeconomic variables that are commonly used in travel demand modeling applications are similar between the targeted marketing data and the household travel survey. Furthermore, for the great majority of MPOs that continue to maintain aggregate four-step travel demand models (which utilize simple medians or means), the targeted marketing data should perform as well as household travel survey data with a significantly larger sample size (0.5% versus an estimated 87.5%).

6.2.2 Household-Level Validation for Hard-to-Reach Groups

This study provided a unique opportunity to look at the worst-case scenario of data accuracy with targeted marketing data by surveying primarily low income, minority neighborhoods and comparing the self-reported sociodemographic information to the targeted marketing

data. It showed that the rate of accuracy between targeted marketing and self-reported data for neighborhoods with hard-to-reach population groups is relatively high for age, gender, and tenure (ranging from 82.3% to 94.5%), but for educational attainment, ethnicity, household income, marital status, number of adults, and number of children in the household, it ranges from 17.4% to 69.0%. The self-reported data also showed that incorrect targeted marketing data randomly occur across all populations in relation to age, gender, household income, number of adults in the household, and tenure. It does not randomly occur across ethnicity or marital status groups. Educational attainment and the number of children in the household were not testable with regards to randomness across groups.

Household-level targeted marketing data was purchased at a low cost for 6,554 households, which was 100% of the households that the targeted marketing firm had data for in the selected neighborhoods. Due to the high costs of surveying, a random sample of only 2,000 of these households could be contacted. Only 5.8% of these 2,000 households responded (n=116). Because the number of respondents to the survey was particularly low, the results from this study may not be representative of the actual accuracy rate for hard-to-reach population groups. Additionally, it is estimated that the self-reported survey data had a significant nonresponse bias (as is the case for any survey with response rates lower than about 90%). This fact emphasizes the difficulty associated with survey-based data collection today.

In future research, a household-level comparison should be conducted for a random sample of households rather than for neighborhoods with hard-to-reach populations. A valuable study would repeat the analysis of Chapter 3 for a larger stratified sample over income and area type (e.g., central business district, urban residential, urban commercial, suburban residential, suburban commercial, exurban, rural).

6.2.3 Airport Passenger Model

Regression models predict the number of home-based trips to the airport from each traffic analysis zone throughout the Atlanta region. Those models based on lifestyle clusters from the targeted marketing data exhibited much higher adjusted R-squared values than did

regression models based solely on income, which is what the current air passenger model for the Atlanta region uses. With the addition of just five of the 26 lifestyle clusters, the R-squared value increased by 67% from 0.2773 to 0.4635. Ultimately, this simple application demonstrated the predictive power of the behavioral and consumer preferences available in the targeted marketing data.

Future research should consider other ways to utilize targeted marketing data to model air passenger trips to and from the airport. The simple regression model used in this application aimed to replicate current practice, but it is anticipated that by using other additional information available for purchase from a targeted marketing firm, an even better model fit could be achieved. Obviously other future research could investigate different trip types as well, including for example educational trips or trips related to special events like professional sports.

6.2.4 Residential Location Choice Model

In this study, discrete choice multinomial logit (MNL) models predict residential location choice behavior for the Atlanta region. This study demonstrated that targeted marketing data can provide new behavioral insights through the large number of variables available. By adding just the lifestyle similarity and dissimilarity measures, the model fit improved dramatically and the significance of these particular variables surpassed any others in the model. This study also emphasized the importance of large sample sizes in specifying robust models through a Monte Carlo experiment. The effect of population sample size on model variance is so large that research focused on determining optimal sample sizes for the number of alternatives becomes obsolete. Simply stated, increasing the number of sampled households in a residential choice model improves consistency in model parameter estimates much faster than increasing the number of alternatives in the sample. This highlights one of the key advantages targeted marketing data provides over traditional surveys— larger, less expensive datasets with new analysis variables that are powerful predictors of travel behaviors.

In future research, residential location choice models that incorporate recent methodological advancements such as endogeneity and spatial correlation should be tested. A valuable study would measure if the effect of the additional variables in targeted marketing data changes with more advanced models. It would also measure if the effect on model variability changes due to household sample size and the number of sampled alternatives with more advanced models.

6.3 Research Limitations

Targeted marketing data most notably lack trip information. For this reason, the use of the data will be relatively straightforward for some applications, such as the ones examined in this work, but will be more challenging for others. Most parts of a full travel demand model require more information than is available in targeted marketing data. Accordingly, this research only reveals a first step in exploring the use of targeted marketing data for representing population characteristics of a region within a travel demand modeling context. Future research must identify techniques for overcoming the lack of trip-making behavior if targeted marketing data will ever be able to fully replace the traditional household travel survey.

6.4 Concluding Thoughts

The validation studies of this work suggest that targeted marketing data are similar to U.S. Census data at small geographic levels for basic demographic and socioeconomic information. The studies also suggest that the existing coverage errors are at least similar, if not lower than, the levels of those in the household travel surveys used today to build travel demand models. However, targeted marketing data have the additional benefit of being very inexpensive with coverage rates of about 87.5%. And for each of the individuals in the targeted marketing data, lifestyle and other behavioral information that are not available in Census data or traditional household travel surveys are readily available. The inclusion of lifestyle and other behavioral information, combined with the ability to easily track individuals over time, provide the opportunity to examine many new research questions. The application studies in this work highlighted these benefits particularly well, including the

additional behavioral information of the lifestyle clusters and the large sample sizes.

Looking ahead, the combination of targeted marketing data with other third-party and non-traditional data could be particularly powerful. It offers tremendous opportunities to enhance, or even transform, existing travel demand modeling systems and data collection practices. Inexpensive, up-to-date, and detailed data available at regular intervals as often as every month or quarter would allow researchers to better study the travel behavior effects of particular economic, climate, or political shocks in addition to transportation infrastructure changes. Particularly with the advent of new performance requirements in MAP-21 that will transform the federal surface transportation program to be more focused on performance outcomes, such detailed information and sensitive modeling capabilities will be highly desirable.

APPENDIX A

MONTE CARLO EXPERIMENT ON RESIDENTIAL LOCATION CHOICE MODEL: LIST OF SUMMARY STATISTICS

Table 29: Monte Carlo Parameter Estimates and Standard Errors

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
log(Average Income)													
500	10	-0.575	-0.345	-0.254	-0.272	-0.183	-0.062	0.141	0.142	0.144	0.144	0.146	0.148
500	50	-0.236	-0.195	-0.150	-0.145	-0.127	-0.009	0.122	0.126	0.127	0.127	0.129	0.134
500	100	-0.304	-0.189	-0.084	-0.101	-0.029	0.089	0.122	0.125	0.126	0.127	0.128	0.133
500	200	-0.221	-0.167	-0.129	-0.087	-0.029	0.102	0.122	0.126	0.126	0.126	0.127	0.130
500	400	-0.291	-0.192	-0.115	-0.109	-0.048	0.126	0.120	0.122	0.124	0.125	0.126	0.134
500	797	-0.347	-0.224	-0.088	-0.123	-0.017	0.049	0.121	0.122	0.123	0.124	0.126	0.128
1,000	10	-0.280	-0.199	-0.178	-0.144	-0.135	0.124	0.097	0.100	0.100	0.101	0.103	0.104
1,000	50	-0.234	-0.188	-0.141	-0.128	-0.060	-0.007	0.087	0.089	0.090	0.090	0.091	0.092
1,000	100	-0.214	-0.142	-0.113	-0.092	-0.057	0.087	0.087	0.088	0.089	0.089	0.091	0.092
1,000	200	-0.221	-0.180	-0.094	-0.099	-0.045	0.031	0.085	0.088	0.088	0.088	0.089	0.091
1,000	400	-0.280	-0.192	-0.105	-0.113	-0.031	0.012	0.085	0.087	0.088	0.088	0.089	0.090
1,000	797	-0.300	-0.220	-0.197	-0.177	-0.131	-0.046	0.085	0.085	0.087	0.087	0.088	0.089
2,500	10	-0.225	-0.159	-0.143	-0.135	-0.119	-0.052	0.063	0.063	0.063	0.063	0.064	0.064
2,500	50	-0.274	-0.155	-0.149	-0.129	-0.069	-0.034	0.056	0.056	0.057	0.057	0.057	0.058
2,500	100	-0.180	-0.169	-0.132	-0.125	-0.087	-0.052	0.055	0.055	0.055	0.056	0.056	0.057
2,500	200	-0.176	-0.149	-0.137	-0.120	-0.090	-0.039	0.055	0.056	0.056	0.056	0.056	0.056
2,500	400	-0.179	-0.151	-0.099	-0.114	-0.082	-0.074	0.054	0.055	0.055	0.055	0.056	0.057

Continued on next page

Table 29: *Continued*

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
2,500	797	-0.147	-0.139	-0.116	-0.118	-0.108	-0.080	0.054	0.055	0.055	0.055	0.056	0.056
5,000	10	-0.239	-0.205	-0.186	-0.179	-0.147	-0.116	0.044	0.044	0.045	0.045	0.045	0.045
5,000	50	-0.172	-0.138	-0.127	-0.121	-0.111	-0.041	0.040	0.040	0.040	0.040	0.041	0.041
5,000	100	-0.169	-0.152	-0.117	-0.121	-0.099	-0.059	0.039	0.039	0.040	0.040	0.040	0.040
5,000	200	-0.183	-0.142	-0.115	-0.125	-0.100	-0.090	0.039	0.039	0.039	0.039	0.040	0.040
5,000	400	-0.162	-0.131	-0.106	-0.111	-0.087	-0.074	0.038	0.039	0.039	0.039	0.039	0.039
5,000	797	-0.216	-0.124	-0.111	-0.122	-0.095	-0.081	0.038	0.039	0.039	0.039	0.039	0.040
10,000	10	-0.184	-0.174	-0.159	-0.161	-0.152	-0.138	0.031	0.032	0.032	0.032	0.032	0.032
10,000	50	-0.156	-0.135	-0.120	-0.123	-0.114	-0.095	0.028	0.028	0.028	0.028	0.028	0.029
10,000	100	-0.155	-0.146	-0.127	-0.129	-0.120	-0.101	0.028	0.028	0.028	0.028	0.028	0.028
10,000	200	-0.179	-0.134	-0.114	-0.118	-0.094	-0.078	0.027	0.028	0.028	0.028	0.028	0.028
10,000	400	-0.140	-0.124	-0.118	-0.118	-0.114	-0.089	0.028	0.028	0.028	0.028	0.028	0.028
10,000	797	-0.206	-0.142	-0.123	-0.130	-0.104	-0.101	0.027	0.027	0.028	0.028	0.028	0.028
50,000	10	-0.170	-0.159	-0.153	-0.154	-0.146	-0.143	0.014	0.014	0.014	0.014	0.014	0.014
50,000	50	-0.143	-0.137	-0.131	-0.126	-0.126	-0.092	0.013	0.013	0.013	0.013	0.013	0.013
50,000	100	-0.132	-0.121	-0.115	-0.116	-0.109	-0.102	0.013	0.013	0.013	0.013	0.013	0.013
50,000	200	-0.141	-0.133	-0.127	-0.125	-0.117	-0.108	0.012	0.012	0.012	0.012	0.012	0.012
50,000	400	-0.145	-0.133	-0.125	-0.128	-0.121	-0.118	0.012	0.012	0.012	0.012	0.012	0.012

Continued on next page

Table 29: *Continued*

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
50,000	797	-0.153	-0.144	-0.137	-0.136	-0.129	-0.115	0.012	0.012	0.012	0.012	0.012	0.012
100,000	10	-0.166	-0.158	-0.156	-0.156	-0.155	-0.141	0.010	0.010	0.010	0.010	0.010	0.010
100,000	50	-0.142	-0.140	-0.134	-0.133	-0.127	-0.123	0.009	0.009	0.009	0.009	0.009	0.009
100,000	100	-0.134	-0.128	-0.125	-0.124	-0.121	-0.115	0.009	0.009	0.009	0.009	0.009	0.009
100,000	200	-0.133	-0.129	-0.127	-0.126	-0.124	-0.118	0.009	0.009	0.009	0.009	0.009	0.009
200,000	10	-0.160	-0.154	-0.150	-0.151	-0.148	-0.146	0.007	0.007	0.007	0.007	0.007	0.007
200,000	50	-0.135	-0.134	-0.133	-0.132	-0.130	-0.124	0.006	0.006	0.006	0.006	0.006	0.006
200,000	100	-0.129	-0.126	-0.124	-0.125	-0.123	-0.121	0.006	0.006	0.006	0.006	0.006	0.006
419,713	10	-0.159	-0.157	-0.155	-0.156	-0.155	-0.154	0.005	0.005	0.005	0.005	0.005	0.005
419,713	50	-0.132	-0.131	-0.131	-0.131	-0.130	-0.129	0.004	0.004	0.004	0.004	0.004	0.004
419,713	100	-0.126	-0.125	-0.125	-0.125	-0.125	-0.124	0.004	0.004	0.004	0.004	0.004	0.004
log(Population Density)													
500	10	-0.131	-0.038	-0.018	-0.017	0.025	0.045	0.066	0.067	0.069	0.068	0.069	0.070
500	50	-0.177	-0.054	-0.027	-0.044	-0.010	0.045	0.059	0.062	0.062	0.062	0.064	0.065
500	100	-0.188	-0.102	-0.043	-0.053	-0.009	0.057	0.058	0.061	0.062	0.062	0.063	0.065
500	200	-0.144	-0.096	-0.074	-0.061	-0.025	0.046	0.059	0.060	0.061	0.061	0.062	0.064
500	400	-0.121	-0.084	-0.051	-0.037	0.001	0.116	0.059	0.060	0.062	0.062	0.062	0.066
500	797	-0.216	-0.087	-0.058	-0.067	-0.034	0.028	0.056	0.060	0.061	0.061	0.062	0.063

Continued on next page

Table 29: *Continued*

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
1,000	10	-0.113	-0.102	-0.066	-0.066	-0.033	-0.017	0.046	0.046	0.047	0.047	0.048	0.048
1,000	50	-0.181	-0.087	-0.072	-0.085	-0.067	-0.035	0.042	0.043	0.044	0.043	0.044	0.044
1,000	100	-0.115	-0.096	-0.064	-0.063	-0.048	0.036	0.042	0.043	0.043	0.043	0.044	0.045
1,000	200	-0.090	-0.086	-0.060	-0.050	-0.022	0.035	0.042	0.043	0.044	0.044	0.044	0.045
1,000	400	-0.100	-0.063	-0.056	-0.045	-0.044	0.032	0.042	0.043	0.043	0.043	0.043	0.045
1,000	797	-0.096	-0.052	-0.046	-0.045	-0.036	0.021	0.042	0.043	0.043	0.043	0.043	0.045
2,500	10	-0.147	-0.102	-0.080	-0.078	-0.065	-0.009	0.029	0.030	0.030	0.030	0.030	0.030
2,500	50	-0.123	-0.082	-0.057	-0.065	-0.045	-0.038	0.027	0.028	0.028	0.028	0.028	0.028
2,500	100	-0.093	-0.087	-0.079	-0.064	-0.057	-0.005	0.027	0.027	0.027	0.027	0.027	0.028
2,500	200	-0.092	-0.071	-0.058	-0.060	-0.048	-0.032	0.027	0.027	0.027	0.027	0.028	0.028
2,500	400	-0.071	-0.049	-0.034	-0.038	-0.028	0.002	0.027	0.027	0.028	0.028	0.028	0.028
2,500	797	-0.089	-0.077	-0.066	-0.062	-0.050	-0.021	0.027	0.027	0.027	0.027	0.027	0.028
5,000	10	-0.103	-0.096	-0.082	-0.083	-0.071	-0.062	0.021	0.021	0.021	0.021	0.021	0.021
5,000	50	-0.110	-0.059	-0.056	-0.058	-0.047	-0.038	0.019	0.020	0.020	0.020	0.020	0.020
5,000	100	-0.109	-0.078	-0.067	-0.068	-0.051	-0.045	0.019	0.019	0.019	0.019	0.019	0.020
5,000	200	-0.090	-0.070	-0.064	-0.066	-0.059	-0.051	0.019	0.019	0.019	0.019	0.019	0.019
5,000	400	-0.091	-0.072	-0.065	-0.064	-0.056	-0.032	0.019	0.019	0.019	0.019	0.019	0.019
5,000	797	-0.083	-0.073	-0.056	-0.057	-0.043	-0.030	0.019	0.019	0.019	0.019	0.019	0.019

Continued on next page

Table 29: *Continued*

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
10,000	10	-0.099	-0.085	-0.077	-0.078	-0.072	-0.054	0.015	0.015	0.015	0.015	0.015	0.015
10,000	50	-0.080	-0.072	-0.066	-0.062	-0.053	-0.039	0.014	0.014	0.014	0.014	0.014	0.014
10,000	100	-0.101	-0.078	-0.074	-0.075	-0.064	-0.053	0.013	0.014	0.014	0.014	0.014	0.014
10,000	200	-0.095	-0.083	-0.074	-0.075	-0.066	-0.061	0.013	0.013	0.014	0.014	0.014	0.014
10,000	400	-0.082	-0.067	-0.061	-0.060	-0.057	-0.034	0.013	0.014	0.014	0.014	0.014	0.014
10,000	797	-0.090	-0.073	-0.057	-0.059	-0.045	-0.037	0.013	0.013	0.014	0.014	0.014	0.014
50,000	10	-0.089	-0.078	-0.074	-0.076	-0.073	-0.067	0.007	0.007	0.007	0.007	0.007	0.007
50,000	50	-0.074	-0.072	-0.071	-0.068	-0.062	-0.056	0.006	0.006	0.006	0.006	0.006	0.006
50,000	100	-0.069	-0.066	-0.065	-0.063	-0.060	-0.052	0.006	0.006	0.006	0.006	0.006	0.006
50,000	200	-0.070	-0.067	-0.065	-0.063	-0.060	-0.052	0.006	0.006	0.006	0.006	0.006	0.006
50,000	400	-0.071	-0.068	-0.063	-0.063	-0.058	-0.053	0.006	0.006	0.006	0.006	0.006	0.006
50,000	797	-0.066	-0.063	-0.060	-0.059	-0.059	-0.048	0.006	0.006	0.006	0.006	0.006	0.006
100,000	10	-0.086	-0.076	-0.074	-0.075	-0.072	-0.068	0.005	0.005	0.005	0.005	0.005	0.005
100,000	50	-0.073	-0.072	-0.070	-0.069	-0.067	-0.062	0.004	0.004	0.004	0.004	0.004	0.004
100,000	100	-0.070	-0.067	-0.065	-0.066	-0.065	-0.060	0.004	0.004	0.004	0.004	0.004	0.004
100,000	200	-0.068	-0.067	-0.064	-0.065	-0.063	-0.060	0.004	0.004	0.004	0.004	0.004	0.004
200,000	10	-0.080	-0.077	-0.076	-0.075	-0.073	-0.070	0.003	0.003	0.003	0.003	0.003	0.003
200,000	50	-0.072	-0.070	-0.068	-0.069	-0.067	-0.065	0.003	0.003	0.003	0.003	0.003	0.003

Continued on next page

Table 29: *Continued*

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
200,000	100	-0.071	-0.070	-0.068	-0.068	-0.067	-0.064	0.003	0.003	0.003	0.003	0.003	0.003
419,713	10	-0.077	-0.076	-0.076	-0.076	-0.075	-0.075	0.002	0.002	0.002	0.002	0.002	0.002
419,713	50	-0.070	-0.069	-0.069	-0.069	-0.069	-0.069	0.002	0.002	0.002	0.002	0.002	0.002
419,713	100	-0.068	-0.067	-0.067	-0.067	-0.067	-0.067	0.002	0.002	0.002	0.002	0.002	0.002
log(Number of Housing Units)													
500	10	1.040	1.100	1.140	1.140	1.160	1.280	0.121	0.123	0.123	0.124	0.124	0.126
500	50	0.900	0.992	1.090	1.100	1.190	1.320	0.110	0.112	0.112	0.112	0.113	0.114
500	100	0.960	1.060	1.140	1.130	1.180	1.290	0.109	0.110	0.111	0.111	0.112	0.112
500	200	0.919	0.995	1.040	1.070	1.160	1.190	0.108	0.109	0.109	0.109	0.109	0.110
500	400	0.975	1.050	1.090	1.100	1.140	1.280	0.107	0.108	0.109	0.109	0.109	0.111
500	797	0.938	1.030	1.160	1.120	1.180	1.240	0.107	0.109	0.109	0.109	0.109	0.110
1,000	10	0.902	0.945	1.080	1.060	1.120	1.260	0.085	0.086	0.086	0.086	0.087	0.087
1,000	50	0.954	1.050	1.070	1.080	1.090	1.170	0.078	0.079	0.079	0.079	0.079	0.079
1,000	100	1.010	1.050	1.070	1.080	1.120	1.160	0.077	0.078	0.078	0.078	0.078	0.079
1,000	200	0.920	1.000	1.060	1.060	1.130	1.150	0.077	0.077	0.077	0.077	0.077	0.078
1,000	400	0.972	1.040	1.080	1.090	1.120	1.200	0.077	0.077	0.077	0.077	0.077	0.077
1,000	797	0.927	1.010	1.020	1.030	1.050	1.140	0.076	0.077	0.077	0.077	0.077	0.077
2,500	10	0.980	1.010	1.050	1.050	1.090	1.140	0.054	0.054	0.054	0.054	0.055	0.055

Continued on next page

Table 29: *Continued*

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
2,500	50	0.990	1.050	1.060	1.070	1.100	1.130	0.050	0.050	0.050	0.050	0.050	0.050
2,500	100	0.917	1.010	1.050	1.050	1.090	1.140	0.049	0.049	0.049	0.049	0.049	0.050
2,500	200	0.979	1.020	1.040	1.050	1.070	1.140	0.049	0.049	0.049	0.049	0.049	0.049
2,500	400	0.961	1.020	1.040	1.050	1.080	1.170	0.049	0.049	0.049	0.049	0.049	0.049
2,500	797	0.996	1.040	1.070	1.070	1.080	1.160	0.048	0.049	0.049	0.049	0.049	0.049
5,000	10	1.010	1.050	1.060	1.060	1.090	1.120	0.038	0.038	0.039	0.039	0.039	0.039
5,000	50	1.040	1.050	1.070	1.070	1.090	1.110	0.035	0.035	0.035	0.035	0.035	0.035
5,000	100	1.000	1.050	1.070	1.080	1.100	1.190	0.035	0.035	0.035	0.035	0.035	0.035
5,000	200	1.000	1.030	1.060	1.060	1.100	1.110	0.035	0.035	0.035	0.035	0.035	0.035
5,000	400	1.020	1.030	1.040	1.050	1.060	1.090	0.034	0.034	0.034	0.034	0.034	0.035
5,000	797	1.010	1.030	1.050	1.060	1.070	1.130	0.034	0.034	0.034	0.034	0.034	0.035
10,000	10	1.030	1.040	1.060	1.070	1.090	1.120	0.027	0.027	0.027	0.027	0.028	0.028
10,000	50	1.020	1.050	1.070	1.070	1.090	1.110	0.025	0.025	0.025	0.025	0.025	0.025
10,000	100	1.030	1.050	1.070	1.070	1.090	1.110	0.025	0.025	0.025	0.025	0.025	0.025
10,000	200	1.020	1.040	1.050	1.050	1.070	1.080	0.024	0.024	0.024	0.024	0.024	0.025
10,000	400	1.060	1.060	1.070	1.070	1.080	1.090	0.024	0.024	0.024	0.024	0.024	0.024
10,000	797	1.020	1.050	1.060	1.070	1.090	1.100	0.024	0.024	0.024	0.024	0.024	0.024
50,000	10	1.050	1.070	1.070	1.070	1.080	1.080	0.012	0.012	0.012	0.012	0.012	0.012

Continued on next page

Table 29: *Continued*

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
50,000	50	1.030	1.070	1.070	1.070	1.070	1.080	0.011	0.011	0.011	0.011	0.011	0.011
50,000	100	1.050	1.060	1.060	1.060	1.070	1.080	0.011	0.011	0.011	0.011	0.011	0.011
50,000	200	1.060	1.070	1.080	1.070	1.080	1.080	0.011	0.011	0.011	0.011	0.011	0.011
50,000	400	1.060	1.060	1.070	1.070	1.070	1.080	0.011	0.011	0.011	0.011	0.011	0.011
50,000	797	1.050	1.060	1.060	1.070	1.070	1.080	0.011	0.011	0.011	0.011	0.011	0.011
100,000	10	1.060	1.070	1.070	1.070	1.070	1.090	0.009	0.009	0.009	0.009	0.009	0.009
100,000	50	1.050	1.070	1.070	1.070	1.070	1.080	0.008	0.008	0.008	0.008	0.008	0.008
100,000	100	1.060	1.060	1.070	1.070	1.070	1.070	0.008	0.008	0.008	0.008	0.008	0.008
100,000	200	1.060	1.060	1.070	1.070	1.070	1.080	0.008	0.008	0.008	0.008	0.008	0.008
200,000	10	1.070	1.070	1.070	1.070	1.070	1.080	0.006	0.006	0.006	0.006	0.006	0.006
200,000	50	1.060	1.060	1.070	1.070	1.070	1.080	0.006	0.006	0.006	0.006	0.006	0.006
200,000	100	1.060	1.060	1.060	1.060	1.070	1.070	0.005	0.005	0.005	0.005	0.005	0.006
419,713	10	1.070	1.070	1.070	1.070	1.070	1.070	0.004	0.004	0.004	0.004	0.004	0.004
419,713	50	1.060	1.070	1.070	1.070	1.070	1.070	0.004	0.004	0.004	0.004	0.004	0.004
419,713	100	1.070	1.070	1.070	1.070	1.070	1.070	0.004	0.004	0.004	0.004	0.004	0.004
Average Commute Time x *log*(Employment Density)													
500	10	-0.433	-0.244	-0.107	-0.154	-0.042	0.063	0.145	0.158	0.180	0.190	0.183	0.297
500	50	-0.510	-0.355	-0.099	-0.171	-0.001	0.069	0.129	0.146	0.179	0.186	0.226	0.262

Continued on next page

Table 29: *Continued*

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
500	100	-0.674	-0.485	-0.305	-0.322	-0.167	-0.065	0.158	0.176	0.220	0.215	0.248	0.286
500	200	-0.541	-0.475	-0.338	-0.272	-0.060	0.107	0.130	0.166	0.225	0.209	0.251	0.271
500	400	-0.633	-0.453	-0.305	-0.274	-0.116	0.106	0.122	0.186	0.211	0.209	0.234	0.287
500	797	-0.554	-0.342	-0.190	-0.223	-0.113	-0.009	0.154	0.173	0.192	0.197	0.219	0.258
1,000	10	-0.384	-0.299	-0.261	-0.231	-0.170	-0.030	0.113	0.135	0.141	0.145	0.162	0.167
1,000	50	-0.426	-0.281	-0.249	-0.223	-0.138	-0.002	0.111	0.136	0.147	0.143	0.152	0.168
1,000	100	-0.421	-0.350	-0.267	-0.251	-0.149	-0.045	0.113	0.130	0.146	0.144	0.163	0.172
1,000	200	-0.572	-0.349	-0.316	-0.310	-0.295	-0.022	0.105	0.146	0.154	0.151	0.155	0.191
1,000	400	-0.774	-0.324	-0.233	-0.251	-0.109	-0.024	0.111	0.120	0.139	0.142	0.155	0.212
1,000	797	-0.902	-0.432	-0.200	-0.331	-0.129	-0.089	0.119	0.127	0.139	0.152	0.166	0.216
2,500	10	-0.310	-0.261	-0.223	-0.182	-0.085	-0.008	0.069	0.084	0.091	0.089	0.095	0.100
2,500	50	-0.387	-0.225	-0.163	-0.162	-0.100	-0.002	0.068	0.078	0.082	0.083	0.090	0.106
2,500	100	-0.243	-0.204	-0.110	-0.127	-0.053	-0.049	0.067	0.072	0.076	0.078	0.086	0.090
2,500	200	-0.318	-0.229	-0.204	-0.209	-0.166	-0.141	0.079	0.082	0.086	0.087	0.090	0.099
2,500	400	-0.336	-0.298	-0.250	-0.257	-0.214	-0.180	0.086	0.088	0.090	0.091	0.094	0.099
2,500	797	-0.406	-0.358	-0.223	-0.255	-0.163	-0.125	0.078	0.082	0.089	0.091	0.099	0.107
5,000	10	-0.296	-0.195	-0.144	-0.157	-0.127	-0.040	0.054	0.061	0.062	0.062	0.064	0.070
5,000	50	-0.276	-0.226	-0.191	-0.197	-0.186	-0.096	0.055	0.062	0.062	0.062	0.064	0.067

Continued on next page

Table 29: *Continued*

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
5,000	100	-0.276	-0.211	-0.195	-0.199	-0.172	-0.121	0.055	0.060	0.061	0.061	0.063	0.066
5,000	200	-0.275	-0.258	-0.221	-0.214	-0.199	-0.069	0.052	0.061	0.062	0.062	0.064	0.067
5,000	400	-0.287	-0.264	-0.253	-0.212	-0.177	-0.027	0.048	0.060	0.064	0.062	0.065	0.066
5,000	797	-0.411	-0.282	-0.262	-0.258	-0.221	-0.136	0.056	0.062	0.065	0.065	0.067	0.075
10,000	10	-0.221	-0.203	-0.192	-0.179	-0.152	-0.111	0.042	0.043	0.044	0.044	0.045	0.047
10,000	50	-0.276	-0.241	-0.181	-0.192	-0.145	-0.116	0.039	0.041	0.042	0.043	0.045	0.048
10,000	100	-0.230	-0.212	-0.194	-0.193	-0.174	-0.153	0.041	0.041	0.043	0.043	0.044	0.045
10,000	200	-0.226	-0.193	-0.181	-0.177	-0.164	-0.101	0.038	0.041	0.042	0.042	0.043	0.044
10,000	400	-0.305	-0.249	-0.201	-0.212	-0.167	-0.141	0.040	0.041	0.043	0.044	0.045	0.048
10,000	797	-0.258	-0.226	-0.211	-0.210	-0.185	-0.176	0.042	0.043	0.043	0.043	0.044	0.046
50,000	10	-0.223	-0.205	-0.191	-0.192	-0.184	-0.149	0.019	0.020	0.020	0.020	0.021	0.021
50,000	50	-0.217	-0.190	-0.182	-0.180	-0.164	-0.153	0.018	0.019	0.019	0.019	0.019	0.020
50,000	100	-0.206	-0.197	-0.188	-0.188	-0.180	-0.167	0.018	0.019	0.019	0.019	0.019	0.019
50,000	200	-0.249	-0.205	-0.188	-0.196	-0.180	-0.167	0.018	0.019	0.019	0.019	0.019	0.020
50,000	400	-0.236	-0.200	-0.187	-0.193	-0.182	-0.160	0.018	0.019	0.019	0.019	0.019	0.020
50,000	797	-0.240	-0.225	-0.202	-0.206	-0.187	-0.177	0.019	0.019	0.019	0.019	0.020	0.020
100,000	10	-0.207	-0.191	-0.182	-0.185	-0.180	-0.169	0.014	0.014	0.014	0.014	0.014	0.014
100,000	50	-0.209	-0.183	-0.181	-0.181	-0.173	-0.165	0.013	0.013	0.013	0.013	0.013	0.014

Continued on next page

Table 29: *Continued*

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
100,000	100	-0.204	-0.192	-0.181	-0.181	-0.172	-0.156	0.013	0.013	0.013	0.013	0.013	0.014
100,000	200	-0.198	-0.195	-0.189	-0.188	-0.180	-0.174	0.013	0.013	0.013	0.013	0.013	0.014
200,000	10	-0.202	-0.200	-0.190	-0.191	-0.183	-0.180	0.010	0.010	0.010	0.010	0.010	0.010
200,000	50	-0.197	-0.190	-0.185	-0.183	-0.176	-0.169	0.009	0.009	0.009	0.009	0.010	0.010
200,000	100	-0.187	-0.180	-0.178	-0.177	-0.172	-0.167	0.009	0.009	0.009	0.009	0.009	0.009
419,713	10	-0.193	-0.190	-0.188	-0.189	-0.187	-0.186	0.007	0.007	0.007	0.007	0.007	0.007
419,713	50	-0.183	-0.183	-0.182	-0.182	-0.182	-0.181	0.007	0.007	0.007	0.007	0.007	0.007
419,713	100	-0.183	-0.183	-0.182	-0.182	-0.182	-0.180	0.006	0.006	0.006	0.006	0.007	0.007
Household Size x Average Household Size													
500	10	0.120	0.805	1.240	1.130	1.560	1.700	0.664	0.686	0.698	0.699	0.716	0.736
500	50	0.454	0.560	0.734	0.724	0.883	1.010	0.587	0.617	0.622	0.620	0.628	0.638
500	100	-0.058	0.183	0.416	0.474	0.790	1.170	0.586	0.597	0.610	0.611	0.622	0.639
500	200	-0.021	0.120	0.577	0.569	0.761	1.610	0.583	0.602	0.609	0.608	0.620	0.624
500	400	0.260	0.664	1.010	0.930	1.270	1.370	0.577	0.588	0.601	0.598	0.611	0.614
500	797	0.094	0.711	0.916	0.913	1.240	1.350	0.575	0.603	0.614	0.611	0.618	0.637
1,000	10	-0.346	0.024	0.239	0.330	0.602	1.350	0.462	0.475	0.478	0.482	0.493	0.505
1,000	50	0.301	0.654	0.699	0.802	0.807	1.580	0.430	0.436	0.441	0.443	0.448	0.459
1,000	100	0.163	0.454	0.555	0.709	0.992	1.660	0.413	0.420	0.428	0.428	0.437	0.441

Continued on next page

Table 29: *Continued*

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
1,000	200	0.193	0.483	0.605	0.629	0.798	1.090	0.414	0.430	0.432	0.432	0.436	0.443
1,000	400	0.079	0.369	0.746	0.705	0.943	1.320	0.415	0.421	0.424	0.424	0.426	0.433
1,000	797	0.161	0.505	0.880	0.856	1.190	1.550	0.420	0.421	0.426	0.428	0.433	0.441
2,500	10	0.379	0.604	0.690	0.724	0.850	1.100	0.284	0.304	0.305	0.304	0.308	0.309
2,500	50	0.508	0.668	0.726	0.734	0.819	0.956	0.270	0.274	0.276	0.276	0.276	0.282
2,500	100	0.298	0.626	0.718	0.714	0.790	1.080	0.269	0.271	0.272	0.273	0.274	0.278
2,500	200	0.333	0.627	0.810	0.758	0.939	1.060	0.267	0.269	0.271	0.271	0.274	0.276
2,500	400	0.663	0.723	0.769	0.819	0.891	1.130	0.265	0.268	0.270	0.270	0.272	0.277
2,500	797	0.215	0.501	0.701	0.681	0.861	1.150	0.264	0.269	0.269	0.269	0.271	0.272
5,000	10	0.520	0.657	0.703	0.700	0.748	0.915	0.212	0.213	0.213	0.214	0.216	0.216
5,000	50	0.594	0.681	0.791	0.793	0.906	1.030	0.193	0.194	0.195	0.196	0.198	0.200
5,000	100	0.480	0.644	0.672	0.718	0.824	1.040	0.191	0.192	0.192	0.193	0.193	0.196
5,000	200	0.575	0.647	0.689	0.694	0.760	0.828	0.189	0.191	0.192	0.192	0.194	0.195
5,000	400	0.409	0.583	0.838	0.790	0.900	1.180	0.189	0.190	0.191	0.191	0.192	0.194
5,000	797	0.476	0.614	0.753	0.748	0.900	0.972	0.189	0.190	0.191	0.191	0.192	0.195
10,000	10	0.468	0.541	0.675	0.640	0.723	0.786	0.151	0.151	0.152	0.152	0.152	0.154
10,000	50	0.543	0.622	0.669	0.699	0.764	0.899	0.137	0.138	0.139	0.139	0.140	0.140
10,000	100	0.549	0.586	0.694	0.676	0.740	0.861	0.135	0.136	0.137	0.137	0.137	0.138

Continued on next page

Table 29: *Continued*

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
10,000	200	0.520	0.690	0.720	0.729	0.780	0.882	0.134	0.135	0.135	0.135	0.136	0.136
10,000	400	0.660	0.699	0.794	0.783	0.823	0.932	0.134	0.134	0.135	0.135	0.136	0.137
10,000	797	0.558	0.688	0.827	0.791	0.887	0.968	0.134	0.135	0.135	0.135	0.136	0.136
50,000	10	0.673	0.680	0.719	0.729	0.765	0.810	0.068	0.068	0.068	0.068	0.068	0.069
50,000	50	0.645	0.699	0.731	0.753	0.821	0.857	0.061	0.062	0.062	0.062	0.062	0.062
50,000	100	0.706	0.750	0.781	0.769	0.792	0.807	0.060	0.061	0.061	0.061	0.061	0.061
50,000	200	0.641	0.709	0.749	0.741	0.780	0.836	0.060	0.061	0.061	0.061	0.061	0.061
50,000	400	0.686	0.726	0.781	0.767	0.795	0.857	0.060	0.060	0.060	0.060	0.061	0.061
50,000	797	0.729	0.760	0.795	0.795	0.811	0.924	0.060	0.060	0.060	0.060	0.061	0.061
100,000	10	0.645	0.701	0.724	0.724	0.749	0.788	0.048	0.048	0.048	0.048	0.048	0.048
100,000	50	0.695	0.708	0.725	0.735	0.766	0.782	0.044	0.044	0.044	0.044	0.044	0.044
100,000	100	0.706	0.730	0.736	0.736	0.743	0.762	0.043	0.043	0.043	0.043	0.043	0.043
100,000	200	0.682	0.730	0.758	0.753	0.779	0.803	0.043	0.043	0.043	0.043	0.043	0.043
200,000	10	0.679	0.698	0.727	0.723	0.738	0.779	0.034	0.034	0.034	0.034	0.034	0.034
200,000	50	0.699	0.713	0.734	0.739	0.762	0.795	0.031	0.031	0.031	0.031	0.031	0.031
200,000	100	0.686	0.719	0.729	0.727	0.738	0.762	0.030	0.030	0.030	0.030	0.030	0.030
419,713	10	0.719	0.722	0.725	0.728	0.737	0.739	0.023	0.023	0.023	0.023	0.024	0.024
419,713	50	0.721	0.727	0.731	0.730	0.732	0.736	0.021	0.021	0.021	0.021	0.021	0.021

Continued on next page

Table 29: *Continued*

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
419,713	100	0.725	0.729	0.731	0.731	0.732	0.736	0.021	0.021	0.021	0.021	0.021	0.021
Household Income x Average Current Market Value													
500	10	0.092	0.114	0.181	0.175	0.226	0.268	0.058	0.061	0.063	0.063	0.066	0.068
500	50	0.088	0.108	0.116	0.124	0.138	0.183	0.044	0.045	0.053	0.052	0.056	0.060
500	100	0.018	0.070	0.102	0.109	0.157	0.214	0.042	0.044	0.047	0.047	0.048	0.052
500	200	0.028	0.096	0.102	0.111	0.134	0.203	0.043	0.046	0.047	0.047	0.049	0.050
500	400	0.034	0.091	0.106	0.123	0.158	0.211	0.038	0.044	0.046	0.045	0.048	0.050
500	797	0.106	0.118	0.131	0.140	0.164	0.184	0.039	0.042	0.045	0.045	0.047	0.050
1,000	10	-0.009	0.073	0.121	0.115	0.166	0.197	0.040	0.042	0.044	0.044	0.045	0.049
1,000	50	0.073	0.108	0.115	0.116	0.128	0.153	0.033	0.035	0.035	0.035	0.035	0.036
1,000	100	0.071	0.080	0.092	0.095	0.105	0.132	0.032	0.032	0.033	0.033	0.034	0.035
1,000	200	0.047	0.090	0.100	0.111	0.147	0.162	0.029	0.031	0.033	0.033	0.033	0.036
1,000	400	0.085	0.096	0.101	0.106	0.113	0.132	0.029	0.032	0.032	0.032	0.032	0.034
1,000	797	0.091	0.119	0.132	0.134	0.150	0.174	0.030	0.030	0.031	0.031	0.033	0.034
2,500	10	0.119	0.135	0.159	0.152	0.169	0.177	0.027	0.028	0.028	0.028	0.029	0.029
2,500	50	0.091	0.123	0.138	0.134	0.145	0.167	0.021	0.022	0.022	0.022	0.022	0.023
2,500	100	0.104	0.111	0.120	0.120	0.127	0.140	0.020	0.021	0.021	0.021	0.021	0.021
2,500	200	0.087	0.110	0.119	0.118	0.131	0.141	0.019	0.020	0.020	0.020	0.021	0.022

Continued on next page

Table 29: *Continued*

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
2,500	400	0.076	0.115	0.126	0.123	0.133	0.153	0.019	0.019	0.020	0.020	0.021	0.022
2,500	797	0.103	0.118	0.120	0.121	0.124	0.143	0.019	0.019	0.020	0.020	0.020	0.021
5,000	10	0.112	0.139	0.147	0.145	0.152	0.174	0.019	0.019	0.020	0.020	0.020	0.021
5,000	50	0.098	0.116	0.120	0.122	0.132	0.144	0.015	0.015	0.015	0.015	0.016	0.016
5,000	100	0.098	0.112	0.120	0.123	0.132	0.165	0.014	0.014	0.015	0.015	0.015	0.015
5,000	200	0.091	0.114	0.118	0.116	0.124	0.127	0.013	0.014	0.014	0.014	0.014	0.015
5,000	400	0.103	0.113	0.118	0.120	0.127	0.141	0.013	0.014	0.014	0.014	0.014	0.014
5,000	797	0.103	0.116	0.121	0.121	0.125	0.143	0.013	0.014	0.014	0.014	0.014	0.015
10,000	10	0.125	0.141	0.145	0.144	0.153	0.162	0.014	0.014	0.014	0.014	0.014	0.014
10,000	50	0.115	0.119	0.123	0.123	0.127	0.129	0.011	0.011	0.011	0.011	0.011	0.011
10,000	100	0.113	0.124	0.126	0.126	0.128	0.139	0.010	0.010	0.010	0.010	0.010	0.010
10,000	200	0.098	0.118	0.122	0.119	0.123	0.135	0.010	0.010	0.010	0.010	0.010	0.011
10,000	400	0.112	0.117	0.120	0.122	0.127	0.135	0.010	0.010	0.010	0.010	0.010	0.010
10,000	797	0.113	0.120	0.125	0.127	0.131	0.153	0.010	0.010	0.010	0.010	0.010	0.010
50,000	10	0.142	0.143	0.147	0.147	0.149	0.151	0.006	0.006	0.006	0.006	0.006	0.006
50,000	50	0.117	0.122	0.125	0.125	0.128	0.133	0.005	0.005	0.005	0.005	0.005	0.005
50,000	100	0.116	0.118	0.120	0.120	0.122	0.127	0.005	0.005	0.005	0.005	0.005	0.005
50,000	200	0.113	0.118	0.123	0.121	0.124	0.125	0.004	0.004	0.004	0.004	0.005	0.005

Continued on next page

Table 29: *Continued*

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
50,000	400	0.114	0.121	0.122	0.122	0.123	0.129	0.004	0.004	0.004	0.004	0.004	0.004
50,000	797	0.121	0.123	0.124	0.124	0.125	0.129	0.004	0.004	0.004	0.004	0.004	0.004
100,000	10	0.135	0.144	0.146	0.145	0.148	0.150	0.004	0.004	0.004	0.004	0.004	0.004
100,000	50	0.119	0.125	0.126	0.126	0.127	0.130	0.003	0.003	0.003	0.003	0.003	0.003
100,000	100	0.117	0.121	0.123	0.122	0.124	0.125	0.003	0.003	0.003	0.003	0.003	0.003
100,000	200	0.115	0.120	0.121	0.121	0.122	0.123	0.003	0.003	0.003	0.003	0.003	0.003
200,000	10	0.140	0.143	0.145	0.144	0.145	0.148	0.003	0.003	0.003	0.003	0.003	0.003
200,000	50	0.125	0.125	0.126	0.127	0.129	0.130	0.002	0.002	0.002	0.002	0.002	0.002
200,000	100	0.120	0.121	0.122	0.122	0.122	0.124	0.002	0.002	0.002	0.002	0.002	0.002
419,713	10	0.146	0.146	0.147	0.147	0.148	0.148	0.002	0.002	0.002	0.002	0.002	0.002
419,713	50	0.125	0.126	0.126	0.126	0.126	0.127	0.002	0.002	0.002	0.002	0.002	0.002
419,713	100	0.121	0.122	0.122	0.122	0.122	0.123	0.002	0.002	0.002	0.002	0.002	0.002
Lifestyle Similarity Measure													
500	10	0.502	0.512	0.545	0.559	0.600	0.639	0.053	0.054	0.054	0.055	0.057	0.058
500	50	0.452	0.464	0.486	0.489	0.497	0.556	0.040	0.042	0.043	0.043	0.044	0.044
500	100	0.443	0.463	0.484	0.491	0.506	0.561	0.040	0.041	0.042	0.042	0.043	0.045
500	200	0.421	0.446	0.475	0.467	0.488	0.502	0.039	0.039	0.041	0.041	0.041	0.043
500	400	0.378	0.421	0.435	0.436	0.459	0.469	0.038	0.039	0.040	0.040	0.040	0.041

Continued on next page

Table 29: *Continued*

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
500	797	0.380	0.409	0.427	0.428	0.442	0.485	0.035	0.038	0.039	0.039	0.040	0.041
1,000	10	0.510	0.531	0.545	0.559	0.596	0.610	0.037	0.038	0.039	0.039	0.039	0.040
1,000	50	0.439	0.476	0.497	0.489	0.511	0.515	0.030	0.031	0.031	0.031	0.031	0.032
1,000	100	0.443	0.466	0.481	0.484	0.489	0.548	0.028	0.029	0.029	0.029	0.029	0.031
1,000	200	0.430	0.452	0.462	0.459	0.468	0.480	0.027	0.028	0.028	0.028	0.028	0.029
1,000	400	0.414	0.427	0.446	0.449	0.469	0.485	0.027	0.027	0.028	0.028	0.028	0.030
1,000	797	0.372	0.382	0.413	0.407	0.427	0.442	0.024	0.026	0.026	0.026	0.027	0.028
2,500	10	0.492	0.517	0.535	0.533	0.555	0.559	0.024	0.024	0.024	0.024	0.024	0.025
2,500	50	0.463	0.467	0.484	0.483	0.490	0.514	0.019	0.019	0.019	0.019	0.019	0.020
2,500	100	0.442	0.458	0.477	0.473	0.490	0.496	0.018	0.018	0.018	0.018	0.019	0.019
2,500	200	0.450	0.456	0.463	0.463	0.470	0.477	0.017	0.018	0.018	0.018	0.018	0.018
2,500	400	0.415	0.431	0.447	0.442	0.453	0.462	0.017	0.017	0.018	0.018	0.018	0.018
2,500	797	0.405	0.422	0.434	0.431	0.442	0.451	0.017	0.017	0.017	0.017	0.018	0.018
5,000	10	0.499	0.537	0.542	0.541	0.552	0.560	0.017	0.017	0.017	0.017	0.017	0.017
5,000	50	0.473	0.475	0.484	0.485	0.493	0.511	0.013	0.014	0.014	0.014	0.014	0.014
5,000	100	0.436	0.463	0.482	0.473	0.484	0.488	0.013	0.013	0.013	0.013	0.013	0.013
5,000	200	0.441	0.453	0.458	0.459	0.464	0.487	0.012	0.013	0.013	0.013	0.013	0.013
5,000	400	0.433	0.442	0.450	0.452	0.459	0.473	0.012	0.012	0.013	0.013	0.013	0.013

Continued on next page

Table 29: *Continued*

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
5,000	797	0.405	0.427	0.436	0.435	0.444	0.460	0.012	0.012	0.012	0.012	0.012	0.012
10,000	10	0.529	0.534	0.545	0.543	0.546	0.562	0.012	0.012	0.012	0.012	0.012	0.012
10,000	50	0.474	0.485	0.491	0.490	0.495	0.506	0.010	0.010	0.010	0.010	0.010	0.010
10,000	100	0.468	0.472	0.475	0.476	0.479	0.489	0.009	0.009	0.009	0.009	0.009	0.009
10,000	200	0.438	0.452	0.458	0.459	0.469	0.473	0.009	0.009	0.009	0.009	0.009	0.009
10,000	400	0.445	0.447	0.451	0.452	0.454	0.463	0.009	0.009	0.009	0.009	0.009	0.009
10,000	797	0.416	0.425	0.436	0.434	0.442	0.449	0.009	0.009	0.009	0.009	0.009	0.009
50,000	10	0.534	0.539	0.541	0.541	0.544	0.550	0.005	0.005	0.005	0.005	0.005	0.005
50,000	50	0.480	0.483	0.488	0.487	0.489	0.494	0.004	0.004	0.004	0.004	0.004	0.004
50,000	100	0.469	0.471	0.472	0.471	0.472	0.473	0.004	0.004	0.004	0.004	0.004	0.004
50,000	200	0.453	0.458	0.464	0.462	0.464	0.468	0.004	0.004	0.004	0.004	0.004	0.004
50,000	400	0.444	0.448	0.450	0.449	0.451	0.452	0.004	0.004	0.004	0.004	0.004	0.004
50,000	797	0.425	0.432	0.434	0.433	0.435	0.439	0.004	0.004	0.004	0.004	0.004	0.004
100,000	10	0.540	0.542	0.543	0.543	0.544	0.546	0.004	0.004	0.004	0.004	0.004	0.004
100,000	50	0.486	0.487	0.488	0.488	0.490	0.492	0.003	0.003	0.003	0.003	0.003	0.003
100,000	100	0.469	0.470	0.472	0.472	0.474	0.476	0.003	0.003	0.003	0.003	0.003	0.003
100,000	200	0.460	0.461	0.464	0.463	0.464	0.465	0.003	0.003	0.003	0.003	0.003	0.003
200,000	10	0.539	0.541	0.543	0.543	0.544	0.547	0.003	0.003	0.003	0.003	0.003	0.003

Continued on next page

Table 29: *Continued*

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
200,000	50	0.483	0.486	0.488	0.487	0.488	0.490	0.002	0.002	0.002	0.002	0.002	0.002
200,000	100	0.472	0.473	0.473	0.473	0.474	0.476	0.002	0.002	0.002	0.002	0.002	0.002
419,713	10	0.541	0.542	0.542	0.542	0.543	0.544	0.002	0.002	0.002	0.002	0.002	0.002
419,713	50	0.486	0.486	0.486	0.487	0.487	0.488	0.002	0.002	0.002	0.002	0.002	0.002
419,713	100	0.472	0.473	0.473	0.473	0.474	0.474	0.001	0.001	0.001	0.001	0.001	0.001
Lifestyle Dissimilarity Measure													
500	10	-0.491	-0.338	-0.307	-0.318	-0.265	-0.208	0.090	0.092	0.094	0.095	0.097	0.101
500	50	-0.563	-0.430	-0.386	-0.384	-0.322	-0.248	0.084	0.086	0.088	0.088	0.090	0.097
500	100	-0.443	-0.413	-0.349	-0.361	-0.322	-0.239	0.083	0.085	0.088	0.087	0.088	0.090
500	200	-0.484	-0.385	-0.354	-0.358	-0.326	-0.235	0.079	0.085	0.086	0.085	0.086	0.090
500	400	-0.556	-0.450	-0.407	-0.414	-0.366	-0.267	0.081	0.085	0.086	0.086	0.088	0.093
500	797	-0.545	-0.411	-0.387	-0.395	-0.381	-0.294	0.079	0.083	0.084	0.085	0.086	0.094
1,000	10	-0.451	-0.392	-0.317	-0.342	-0.297	-0.269	0.065	0.066	0.068	0.067	0.068	0.069
1,000	50	-0.489	-0.388	-0.375	-0.367	-0.334	-0.288	0.060	0.061	0.062	0.062	0.063	0.064
1,000	100	-0.510	-0.392	-0.374	-0.376	-0.329	-0.316	0.060	0.061	0.061	0.061	0.062	0.064
1,000	200	-0.500	-0.442	-0.416	-0.420	-0.392	-0.359	0.060	0.061	0.061	0.062	0.062	0.065
1,000	400	-0.478	-0.449	-0.416	-0.414	-0.393	-0.341	0.058	0.060	0.061	0.061	0.062	0.063
1,000	797	-0.531	-0.496	-0.466	-0.463	-0.421	-0.397	0.058	0.060	0.061	0.061	0.061	0.063

Continued on next page

Table 29: *Continued*

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
2,500	10	-0.373	-0.356	-0.336	-0.333	-0.307	-0.290	0.041	0.041	0.042	0.042	0.042	0.043
2,500	50	-0.410	-0.367	-0.360	-0.361	-0.351	-0.317	0.038	0.039	0.039	0.039	0.039	0.040
2,500	100	-0.410	-0.380	-0.364	-0.369	-0.358	-0.345	0.038	0.038	0.038	0.038	0.039	0.040
2,500	200	-0.432	-0.426	-0.406	-0.404	-0.386	-0.371	0.038	0.039	0.039	0.039	0.039	0.040
2,500	400	-0.505	-0.446	-0.427	-0.433	-0.416	-0.386	0.038	0.038	0.039	0.039	0.039	0.040
2,500	797	-0.517	-0.447	-0.418	-0.431	-0.397	-0.384	0.038	0.038	0.038	0.038	0.039	0.040
5,000	10	-0.346	-0.332	-0.324	-0.326	-0.320	-0.305	0.029	0.029	0.030	0.030	0.030	0.030
5,000	50	-0.421	-0.380	-0.369	-0.373	-0.359	-0.348	0.027	0.028	0.028	0.028	0.028	0.029
5,000	100	-0.430	-0.414	-0.399	-0.398	-0.387	-0.354	0.027	0.028	0.028	0.028	0.028	0.029
5,000	200	-0.448	-0.422	-0.414	-0.413	-0.408	-0.364	0.027	0.027	0.028	0.028	0.028	0.028
5,000	400	-0.435	-0.414	-0.404	-0.400	-0.382	-0.362	0.027	0.027	0.027	0.027	0.028	0.028
5,000	797	-0.470	-0.447	-0.436	-0.425	-0.408	-0.360	0.027	0.027	0.027	0.027	0.028	0.028
10,000	10	-0.388	-0.353	-0.338	-0.342	-0.324	-0.323	0.021	0.021	0.021	0.021	0.021	0.022
10,000	50	-0.404	-0.389	-0.381	-0.377	-0.367	-0.336	0.019	0.020	0.020	0.020	0.020	0.020
10,000	100	-0.407	-0.396	-0.390	-0.387	-0.383	-0.358	0.019	0.019	0.019	0.019	0.020	0.020
10,000	200	-0.423	-0.409	-0.400	-0.399	-0.390	-0.365	0.019	0.019	0.019	0.019	0.019	0.020
10,000	400	-0.441	-0.425	-0.405	-0.412	-0.401	-0.390	0.019	0.019	0.019	0.019	0.019	0.020
10,000	797	-0.449	-0.427	-0.422	-0.414	-0.398	-0.376	0.019	0.019	0.019	0.019	0.019	0.019

Continued on next page

Table 29: *Continued*

		Estimate						Standard Error					
Observations	Alternatives	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
50,000	10	-0.357	-0.347	-0.346	-0.343	-0.343	-0.322	0.009	0.009	0.009	0.009	0.009	0.009
50,000	50	-0.392	-0.383	-0.372	-0.374	-0.366	-0.356	0.009	0.009	0.009	0.009	0.009	0.009
50,000	100	-0.397	-0.394	-0.388	-0.389	-0.387	-0.376	0.009	0.009	0.009	0.009	0.009	0.009
50,000	200	-0.413	-0.405	-0.403	-0.400	-0.393	-0.383	0.009	0.009	0.009	0.009	0.009	0.009
50,000	400	-0.416	-0.414	-0.411	-0.411	-0.409	-0.402	0.009	0.009	0.009	0.009	0.009	0.009
50,000	797	-0.433	-0.426	-0.424	-0.423	-0.420	-0.416	0.009	0.009	0.009	0.009	0.009	0.009
100,000	10	-0.343	-0.338	-0.336	-0.337	-0.334	-0.332	0.007	0.007	0.007	0.007	0.007	0.007
100,000	50	-0.389	-0.379	-0.374	-0.375	-0.371	-0.368	0.006	0.006	0.006	0.006	0.006	0.006
100,000	100	-0.399	-0.392	-0.389	-0.390	-0.386	-0.379	0.006	0.006	0.006	0.006	0.006	0.006
100,000	200	-0.410	-0.401	-0.400	-0.399	-0.395	-0.391	0.006	0.006	0.006	0.006	0.006	0.006
200,000	10	-0.342	-0.341	-0.340	-0.339	-0.338	-0.333	0.005	0.005	0.005	0.005	0.005	0.005
200,000	50	-0.381	-0.378	-0.376	-0.376	-0.374	-0.372	0.004	0.004	0.004	0.004	0.004	0.004
200,000	100	-0.393	-0.391	-0.388	-0.389	-0.387	-0.384	0.004	0.004	0.004	0.004	0.004	0.004
419,713	10	-0.339	-0.338	-0.338	-0.338	-0.337	-0.335	0.003	0.003	0.003	0.003	0.003	0.003
419,713	50	-0.377	-0.377	-0.377	-0.377	-0.377	-0.376	0.003	0.003	0.003	0.003	0.003	0.003
419,713	100	-0.389	-0.388	-0.388	-0.388	-0.388	-0.387	0.003	0.003	0.003	0.003	0.003	0.003

APPENDIX B

MONTE CARLO EXPERIMENT ON RESIDENTIAL LOCATION CHOICE MODEL: SUMMARY OF RESULTS BY VARIABLE

Table 30: *log*(Average Income)

	Alts						Alts					
Obs	10	50	100	200	400	797	10	50	100	200	400	797
	Mean of Estimates						Minimum of Estimates					
500	-0.272	-0.145	-0.101	-0.087	-0.109	-0.123	-0.575	-0.236	-0.304	-0.221	-0.291	-0.347
1,000	-0.144	-0.128	-0.092	-0.099	-0.113	-0.177	-0.280	-0.234	-0.214	-0.221	-0.280	-0.300
2,500	-0.135	-0.129	-0.125	-0.120	-0.114	-0.118	-0.225	-0.274	-0.180	-0.176	-0.179	-0.147
5,000	-0.179	-0.121	-0.121	-0.125	-0.111	-0.122	-0.239	-0.172	-0.169	-0.183	-0.162	-0.216
10,000	-0.161	-0.123	-0.129	-0.118	-0.118	-0.130	-0.184	-0.156	-0.155	-0.179	-0.140	-0.206
50,000	-0.154	-0.126	-0.116	-0.125	-0.128	-0.136	-0.170	-0.143	-0.132	-0.141	-0.145	-0.153
100,000	-0.156	-0.133	-0.124	-0.126	—	—	-0.166	-0.142	-0.134	-0.133	—	—
200,000	-0.151	-0.132	-0.125	—	—	—	-0.160	-0.135	-0.129	—	—	—
419,713	-0.156	-0.131	-0.125	—	—	—	-0.159	-0.132	-0.126	—	—	—
	Standard Deviation of Estimates						Maximum of Estimates					
500	0.143	0.067	0.138	0.111	0.127	0.135	-0.062	-0.009	0.089	0.102	0.126	0.049
1,000	0.123	0.078	0.085	0.088	0.098	0.080	0.124	-0.007	0.087	0.031	0.012	-0.046
2,500	0.051	0.074	0.049	0.045	0.039	0.023	-0.052	-0.034	-0.052	-0.039	-0.074	-0.080
5,000	0.042	0.040	0.035	0.032	0.030	0.042	-0.116	-0.041	-0.059	-0.090	-0.074	-0.081
10,000	0.016	0.021	0.018	0.031	0.014	0.032	-0.138	-0.095	-0.101	-0.078	-0.089	-0.101
50,000	0.009	0.017	0.010	0.011	0.009	0.012	-0.143	-0.092	-0.102	-0.108	-0.118	-0.115
100,000	0.007	0.007	0.006	0.004	—	—	-0.141	-0.123	-0.115	-0.118	—	—
200,000	0.005	0.004	0.002	—	—	—	-0.146	-0.124	-0.121	—	—	—
419,713	0.002	0.001	0.001	—	—	—	-0.154	-0.129	-0.124	—	—	—

— = Due to limitations of the random access memory (RAM) on the computer used for estimation, these models were not estimated.

Table 31: *log*(Population Density)

	Alts						Alts					
Obs	10	50	100	200	400	797	10	50	100	200	400	797
	Mean of Estimates						Minimum of Estimates					
500	-0.017	-0.044	-0.054	-0.061	-0.037	-0.067	-0.131	-0.177	-0.188	-0.144	-0.121	-0.216
1,000	-0.066	-0.085	-0.063	-0.050	-0.045	-0.045	-0.113	-0.181	-0.115	-0.090	-0.100	-0.096
2,500	-0.078	-0.065	-0.064	-0.060	-0.038	-0.062	-0.147	-0.123	-0.093	-0.092	-0.071	-0.089
5,000	-0.083	-0.058	-0.068	-0.066	-0.064	-0.057	-0.103	-0.110	-0.109	-0.090	-0.091	-0.083
10,000	-0.078	-0.063	-0.075	-0.075	-0.060	-0.059	-0.099	-0.080	-0.101	-0.095	-0.081	-0.090
50,000	-0.076	-0.067	-0.063	-0.063	-0.063	-0.059	-0.089	-0.074	-0.069	-0.070	-0.071	-0.066
100,000	-0.075	-0.069	-0.066	-0.065	—	—	-0.086	-0.073	-0.070	-0.068	—	—
200,000	-0.075	-0.068	-0.068	—	—	—	-0.080	-0.072	-0.071	—	—	—
419,713	-0.076	-0.069	-0.067	—	—	—	-0.077	-0.070	-0.068	—	—	—
	Standard Deviation of Estimates						Maximum of Estimates					
500	0.051	0.069	0.077	0.060	0.073	0.065	0.045	0.045	0.057	0.046	0.116	0.028
1,000	0.039	0.043	0.043	0.042	0.040	0.030	-0.017	-0.035	0.036	0.035	0.032	0.021
2,500	0.041	0.027	0.032	0.021	0.021	0.022	-0.009	-0.038	-0.005	-0.032	0.002	-0.021
5,000	0.015	0.020	0.020	0.011	0.017	0.018	-0.062	-0.038	-0.045	-0.051	-0.032	-0.030
10,000	0.012	0.013	0.016	0.012	0.015	0.018	-0.054	-0.039	-0.053	-0.062	-0.034	-0.037
50,000	0.007	0.007	0.005	0.006	0.006	0.006	-0.067	-0.056	-0.052	-0.052	-0.053	-0.048
100,000	0.005	0.003	0.003	0.003	—	—	-0.068	-0.062	-0.060	-0.061	—	—
200,000	0.003	0.002	0.002	—	—	—	-0.070	-0.065	-0.064	—	—	—
419,713	0.001	0.000	0.000	—	—	—	-0.075	-0.068	-0.067	—	—	—

— = Due to limitations of the random access memory (RAM) on the computer used for estimation, these models were not estimated.

Table 32: *log*(Number of Housing Units)

	Alts						Alts					
Obs	10	50	100	200	400	797	10	50	100	200	400	797
	Mean of Estimates						Minimum of Estimates					
500	1.145	1.099	1.132	1.066	1.104	1.118	1.043	0.900	0.960	0.919	0.975	0.938
1,000	1.055	1.076	1.079	1.060	1.087	1.032	0.902	0.954	1.006	0.920	0.972	0.927
2,500	1.053	1.066	1.047	1.051	1.048	1.067	0.980	0.990	0.917	0.979	0.961	0.996
5,000	1.063	1.073	1.079	1.064	1.047	1.059	1.008	1.037	1.003	1.002	1.020	1.013
10,000	1.069	1.068	1.070	1.055	1.070	1.065	1.029	1.025	1.035	1.020	1.057	1.015
50,000	1.071	1.066	1.064	1.074	1.068	1.066	1.052	1.027	1.046	1.059	1.057	1.055
100,000	1.070	1.068	1.068	1.067	—	—	1.057	1.048	1.064	1.057	—	—
200,000	1.070	1.068	1.064	—	—	—	1.067	1.061	1.058	—	—	—
419,713	1.069	1.066	1.066	—	—	—	1.067	1.065	1.066	—	—	—
	Standard Deviation of Estimates						Maximum of Estimates					
500	0.074	0.138	0.098	0.095	0.087	0.102	1.284	1.323	1.286	1.189	1.279	1.240
1,000	0.117	0.062	0.050	0.078	0.069	0.059	1.256	1.168	1.163	1.148	1.200	1.136
2,500	0.054	0.042	0.061	0.047	0.065	0.048	1.143	1.127	1.137	1.136	1.170	1.163
5,000	0.037	0.024	0.052	0.040	0.023	0.042	1.115	1.113	1.193	1.112	1.094	1.132
10,000	0.035	0.029	0.026	0.021	0.010	0.026	1.123	1.112	1.111	1.082	1.086	1.099
50,000	0.009	0.015	0.009	0.009	0.008	0.009	1.082	1.079	1.078	1.081	1.083	1.083
100,000	0.008	0.008	0.003	0.007	—	—	1.085	1.081	1.072	1.079	—	—
200,000	0.003	0.006	0.004	—	—	—	1.078	1.078	1.069	—	—	—
419,713	0.001	0.001	0.000	—	—	—	1.071	1.067	1.067	—	—	—

— = Due to limitations of the random access memory (RAM) on the computer used for estimation, these models were not estimated.

Table 33: Average Commute Time x *log*(Employment Density)

	Alts						Alts					
Obs	10	50	100	200	400	797	10	50	100	200	400	797
	Mean of Estimates						Minimum of Estimates					
500	-0.154	-0.171	-0.322	-0.272	-0.274	-0.223	-0.433	-0.510	-0.674	-0.541	-0.633	-0.554
1,000	-0.231	-0.223	-0.251	-0.310	-0.251	-0.331	-0.384	-0.426	-0.421	-0.572	-0.774	-0.902
2,500	-0.182	-0.162	-0.127	-0.209	-0.257	-0.255	-0.310	-0.387	-0.243	-0.318	-0.336	-0.406
5,000	-0.157	-0.197	-0.199	-0.214	-0.212	-0.258	-0.296	-0.276	-0.276	-0.275	-0.287	-0.411
10,000	-0.179	-0.192	-0.193	-0.177	-0.212	-0.210	-0.221	-0.276	-0.230	-0.226	-0.305	-0.258
50,000	-0.192	-0.180	-0.188	-0.196	-0.193	-0.206	-0.223	-0.217	-0.206	-0.249	-0.236	-0.240
100,000	-0.185	-0.181	-0.181	-0.188	—	—	-0.207	-0.209	-0.204	-0.198	—	—
200,000	-0.191	-0.183	-0.177	—	—	—	-0.202	-0.197	-0.187	—	—	—
419,713	-0.189	-0.182	-0.182	—	—	—	-0.193	-0.183	-0.183	—	—	—
	Standard Deviation of Estimates						Maximum of Estimates					
500	0.164	0.219	0.210	0.232	0.238	0.173	0.063	0.069	-0.065	0.107	0.106	-0.009
1,000	0.109	0.124	0.132	0.136	0.217	0.280	-0.030	-0.002	-0.045	-0.022	-0.024	-0.089
2,500	0.108	0.117	0.080	0.056	0.053	0.112	-0.008	-0.002	-0.049	-0.141	-0.180	-0.125
5,000	0.073	0.053	0.047	0.061	0.084	0.072	-0.040	-0.096	-0.121	-0.069	-0.027	-0.136
10,000	0.035	0.058	0.026	0.033	0.058	0.027	-0.111	-0.116	-0.153	-0.101	-0.141	-0.176
50,000	0.021	0.019	0.012	0.024	0.023	0.023	-0.149	-0.153	-0.167	-0.167	-0.160	-0.177
100,000	0.012	0.013	0.015	0.009	—	—	-0.169	-0.165	-0.156	-0.174	—	—
200,000	0.009	0.009	0.006	—	—	—	-0.180	-0.169	-0.167	—	—	—
419,713	0.002	0.001	0.001	—	—	—	-0.186	-0.181	-0.180	—	—	—

— = Due to limitations of the random access memory (RAM) on the computer used for estimation, these models were not estimated.

Table 34: Household Size x Average Household Size

	Alts						Alts					
Obs	10	50	100	200	400	797	10	50	100	200	400	797
	Mean of Estimates						Minimum of Estimates					
500	1.131	0.724	0.474	0.569	0.930	0.913	0.120	0.454	-0.058	-0.020	0.260	0.094
1,000	0.330	0.802	0.709	0.629	0.705	0.856	-0.346	0.301	0.163	0.193	0.079	0.161
2,500	0.724	0.734	0.714	0.758	0.819	0.681	0.379	0.508	0.298	0.333	0.663	0.215
5,000	0.700	0.793	0.718	0.694	0.790	0.748	0.520	0.594	0.480	0.575	0.409	0.476
10,000	0.640	0.699	0.676	0.729	0.783	0.791	0.468	0.543	0.549	0.520	0.660	0.558
50,000	0.729	0.753	0.769	0.741	0.767	0.795	0.673	0.645	0.706	0.641	0.686	0.729
100,000	0.724	0.735	0.736	0.753	—	—	0.645	0.695	0.706	0.682	—	—
200,000	0.723	0.739	0.727	—	—	—	0.679	0.699	0.686	—	—	—
419,713	0.728	0.730	0.731	—	—	—	0.719	0.721	0.725	—	—	—
	Standard Deviation of Estimates						Maximum of Estimates					
500	0.526	0.196	0.400	0.528	0.410	0.386	1.701	1.011	1.166	1.611	1.366	1.348
1,000	0.495	0.408	0.460	0.285	0.405	0.438	1.345	1.581	1.658	1.091	1.318	1.546
2,500	0.213	0.133	0.207	0.234	0.140	0.315	1.101	0.956	1.082	1.065	1.126	1.146
5,000	0.120	0.152	0.161	0.082	0.256	0.180	0.915	1.027	1.039	0.828	1.183	0.972
10,000	0.117	0.116	0.104	0.108	0.095	0.130	0.786	0.899	0.861	0.882	0.932	0.968
50,000	0.054	0.075	0.036	0.058	0.053	0.055	0.810	0.857	0.807	0.836	0.857	0.924
100,000	0.042	0.033	0.015	0.035	—	—	0.788	0.782	0.762	0.803	—	—
200,000	0.032	0.032	0.021	—	—	—	0.779	0.795	0.762	—	—	—
419,713	0.008	0.004	0.003	—	—	—	0.739	0.736	0.736	—	—	—

— = Due to limitations of the random access memory (RAM) on the computer used for estimation, these models were not estimated.

Table 35: Household Income x Average Current Market Value

	Alts						Alts					
Obs	10	50	100	200	400	797	10	50	100	200	400	797
	Mean of Estimates						Minimum of Estimates					
500	0.175	0.124	0.109	0.111	0.123	0.140	0.093	0.088	0.017	0.028	0.034	0.106
1,000	0.115	0.116	0.095	0.111	0.106	0.134	-0.009	0.074	0.071	0.047	0.085	0.091
2,500	0.152	0.134	0.120	0.118	0.123	0.121	0.119	0.091	0.104	0.087	0.076	0.103
5,000	0.145	0.122	0.123	0.116	0.120	0.121	0.112	0.098	0.098	0.091	0.103	0.103
10,000	0.144	0.123	0.126	0.119	0.122	0.127	0.125	0.115	0.113	0.098	0.112	0.113
50,000	0.147	0.125	0.120	0.121	0.122	0.124	0.142	0.117	0.116	0.113	0.114	0.121
100,000	0.145	0.126	0.122	0.121	—	—	0.135	0.119	0.117	0.115	—	—
200,000	0.144	0.127	0.122	—	—	—	0.140	0.125	0.120	—	—	—
419,713	0.147	0.126	0.122	—	—	—	0.146	0.125	0.121	—	—	—
	Standard Deviation of Estimates						Maximum of Estimates					
500	0.064	0.029	0.063	0.045	0.056	0.029	0.268	0.183	0.214	0.203	0.211	0.184
1,000	0.063	0.023	0.020	0.037	0.015	0.025	0.197	0.153	0.132	0.162	0.132	0.174
2,500	0.022	0.021	0.012	0.016	0.021	0.012	0.177	0.167	0.140	0.141	0.153	0.143
5,000	0.016	0.015	0.019	0.010	0.011	0.012	0.174	0.144	0.165	0.127	0.141	0.143
10,000	0.012	0.005	0.007	0.012	0.008	0.011	0.162	0.129	0.139	0.135	0.135	0.153
50,000	0.003	0.005	0.003	0.004	0.004	0.002	0.151	0.133	0.127	0.125	0.129	0.129
100,000	0.004	0.003	0.003	0.002	—	—	0.150	0.130	0.125	0.123	—	—
200,000	0.002	0.002	0.001	—	—	—	0.148	0.130	0.124	—	—	—
419,713	0.001	0.001	0.001	—	—	—	0.148	0.127	0.123	—	—	—

— = Due to limitations of the random access memory (RAM) on the computer used for estimation, these models were not estimated.

Table 36: Lifestyle Similarity Measure

	Alts						Alts					
Obs	10	50	100	200	400	797	10	50	100	200	400	797
	Mean of Estimates						Minimum of Estimates					
500	0.559	0.489	0.491	0.467	0.436	0.428	0.502	0.452	0.443	0.421	0.378	0.380
1,000	0.559	0.489	0.484	0.459	0.449	0.407	0.510	0.439	0.443	0.430	0.414	0.372
2,500	0.533	0.483	0.473	0.463	0.442	0.431	0.492	0.463	0.442	0.450	0.415	0.405
5,000	0.541	0.485	0.473	0.459	0.452	0.435	0.499	0.473	0.436	0.441	0.433	0.405
10,000	0.543	0.490	0.476	0.459	0.452	0.434	0.529	0.474	0.468	0.438	0.445	0.416
50,000	0.541	0.487	0.471	0.462	0.449	0.433	0.534	0.480	0.469	0.453	0.444	0.425
100,000	0.543	0.488	0.472	0.463	—	—	0.540	0.486	0.469	0.460	—	—
200,000	0.543	0.487	0.473	—	—	—	0.539	0.483	0.472	—	—	—
419,713	0.542	0.487	0.473	—	—	—	0.541	0.486	0.472	—	—	—
	Standard Deviation of Estimates						Maximum of Estimates					
500	0.054	0.035	0.039	0.030	0.028	0.030	0.639	0.556	0.561	0.502	0.469	0.485
1,000	0.037	0.028	0.029	0.014	0.025	0.025	0.610	0.515	0.548	0.480	0.485	0.442
2,500	0.024	0.018	0.019	0.010	0.015	0.016	0.559	0.514	0.496	0.477	0.462	0.451
5,000	0.017	0.013	0.017	0.013	0.013	0.015	0.560	0.511	0.488	0.487	0.473	0.460
10,000	0.010	0.010	0.006	0.011	0.006	0.012	0.562	0.506	0.489	0.473	0.463	0.449
50,000	0.005	0.004	0.001	0.005	0.002	0.004	0.550	0.494	0.473	0.468	0.452	0.439
100,000	0.002	0.002	0.002	0.002	—	—	0.546	0.492	0.476	0.465	—	—
200,000	0.002	0.002	0.001	—	—	—	0.547	0.490	0.476	—	—	—
419,713	0.001	0.001	0.001	—	—	—	0.544	0.488	0.474	—	—	—

— = Due to limitations of the random access memory (RAM) on the computer used for estimation, these models were not estimated.

Table 37: Lifestyle Dissimilarity Measure

	Alts						Alts					
Obs	10	50	100	200	400	797	10	50	100	200	400	797
	Mean of Estimates						Minimum of Estimates					
500	-0.318	-0.384	-0.361	-0.358	-0.414	-0.395	-0.491	-0.563	-0.443	-0.484	-0.556	-0.545
1,000	-0.342	-0.367	-0.376	-0.420	-0.414	-0.463	-0.451	-0.489	-0.510	-0.500	-0.478	-0.531
2,500	-0.333	-0.361	-0.369	-0.404	-0.433	-0.431	-0.373	-0.410	-0.410	-0.432	-0.505	-0.517
5,000	-0.326	-0.373	-0.398	-0.413	-0.400	-0.425	-0.346	-0.421	-0.430	-0.448	-0.435	-0.470
10,000	-0.342	-0.377	-0.387	-0.399	-0.412	-0.414	-0.388	-0.404	-0.407	-0.423	-0.441	-0.449
50,000	-0.343	-0.374	-0.389	-0.400	-0.411	-0.423	-0.357	-0.392	-0.397	-0.413	-0.416	-0.433
100,000	-0.337	-0.375	-0.390	-0.399	—	—	-0.343	-0.389	-0.399	-0.410	—	—
200,000	-0.339	-0.376	-0.389	—	—	—	-0.342	-0.381	-0.393	—	—	—
419,713	-0.338	-0.377	-0.388	—	—	—	-0.339	-0.377	-0.389	—	—	—
	Standard Deviation of Estimates						Maximum of Estimates					
500	0.087	0.089	0.066	0.066	0.087	0.068	-0.208	-0.248	-0.239	-0.235	-0.267	-0.294
1,000	0.060	0.055	0.058	0.041	0.046	0.048	-0.269	-0.288	-0.316	-0.359	-0.341	-0.397
2,500	0.029	0.024	0.020	0.024	0.032	0.048	-0.290	-0.317	-0.345	-0.371	-0.386	-0.384
5,000	0.013	0.022	0.025	0.023	0.024	0.034	-0.305	-0.348	-0.354	-0.364	-0.362	-0.360
10,000	0.021	0.019	0.015	0.018	0.018	0.024	-0.323	-0.336	-0.358	-0.365	-0.390	-0.376
50,000	0.010	0.011	0.006	0.009	0.005	0.005	-0.322	-0.356	-0.376	-0.383	-0.402	-0.416
100,000	0.003	0.007	0.006	0.006	—	—	-0.332	-0.368	-0.379	-0.391	—	—
200,000	0.003	0.003	0.003	—	—	—	-0.333	-0.372	-0.384	—	—	—
419,713	0.001	0.000	0.001	—	—	—	-0.335	-0.376	-0.387	—	—	—

— = Due to limitations of the random access memory (RAM) on the computer used for estimation, these models were not estimated.